

Optimizations of protein force fields

Yoshitake Sakae and Yuko Okamoto

To be published in *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes – from Bioinformatics to Molecular Quantum Mechanics*, edited by Adam Liwo, (Springer, Berlin, 2012).

Abstract

In this Chapter we review our works on force fields for molecular simulations of protein systems. We first discuss the functional forms of the force fields and present some extensions of the conventional ones. We then present various methods for force-field parameter optimizations. Finally, some examples of our applications of these parameter optimization methods are given and they are compared with the results from the existing force-fields.

Yoshitake Sakae

Department of Theoretical and Computational Molecular Science, Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan

Department of Physics, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan

e-mail: sakae@tb.phys.nagoya-u.ac.jp

Yuko Okamoto

Department of Physics, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan

Structural Biology Research Center, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan

Center for Computational Science, Graduate School of Engineering, Nagoya University, Nagoya, Aichi 464-8603, Japan

Information Technology Center, Nagoya University, Nagoya, Aichi 464-8601, Japan

e-mail: okamoto@phys.nagoya-u.ac.jp

1 Introduction

Computer simulations of protein folding into native structures can be achieved when both of the following two requirements are met: (1) potential energy functions (or, force fields) for the protein systems are sufficiently accurate and (2) sufficiently powerful conformational sampling methods are available. Professor Harold A. Scheraga has been one of the most important pioneers in studies of both of the above requirements [1, 2]. By the developments of the generalized-ensemble algorithms (for reviews, see, e.g., Refs. [3, 4, 5, 6]) and related methods, Requirement (2) seems to be almost fulfilled. In this Chapter, we therefore concentrate our attention on Requirement (1).

There are several well-known all-atom (or united-atom) force fields, such as AMBER [7, 8, 9, 10, 11], CHARMM [12, 13, 14], OPLS [15, 16], GROMOS [17, 18], GROMACS [19, 20], and ECEPP [21, 22]. Generally, the force-field parameters are determined based on experimental results for small molecules and theoretical results using quantum chemistry calculations of small peptides such as alanine dipeptide.

However, the simulations using different force-field parameters will give different results. We have performed detailed comparisons of three version of AMBER (ff94 [7], ff96 [8], and ff99 [9]), CHARMM [12], OPLS-AA/L [16], and GROMOS [17] by generalized-ensemble simulations of two small peptides in explicit solvent. [23, 24] We saw that these force fields showed clearly different behaviors especially with respect to secondary-structure-forming tendencies. The folding simulations of the two peptides with implicit solvent model also showed similar results [25, 26, 27]. For instance, the ff94 [7] and ff96 [8] versions of AMBER yield very different behaviors about the secondary-structure-forming tendencies, although these force fields differ only in the main-chain torsion-energy terms. Many researchers have thus studied the main-chain torsion-energy terms and their force-field parameters. For example, newer force-field parameters for the main-chain torsion-energy terms about ϕ and ψ angles have been developed, which are, e.g., AMBER ff99SB [10], AMBER ff03 [11], CHARMM22/CMAP [13, 14] and OPLS-AA/L [16]. The methods of the force-field optimization thus mainly concentrate on the torsion-energy terms. These modifications of the torsion energy are usually based on quantum chemistry calculations [28, 29, 30, 13, 14, 31] or NMR experimental results [32, 33].

We have proposed a new main-chain torsion-energy term, which is represented by a double Fourier series in two variables, the main-chain dihedral angles ϕ and ψ [34, 35]. This expression gives a natural representation of the torsion energy in the Ramachandran space [36] in the sense that any two-dimensional energy surface periodic in both ϕ and ψ can be expanded by the double Fourier series. We can then easily control secondary-structure-forming tendencies by modifying the main-chain torsion-energy surface. We have presented preliminary results for AMBER ff94 and AMBER ff96 [34, 35].

Moreover, we have introduced several optimization methods of force-field parameters [25, 26, 27, 37, 38]. These methods are based on the minimization of some score functions by simulations in the force-field parameter space, where the score functions are derived from the protein coordinate data in the Protein Data Bank

(PDB). One of the score functions consists of the sum of the square of the force acting on each atom in the proteins with the structures from the PDB [25, 26, 27]. Other score functions are taken from the root-mean-square deviations between the original PDB structures and the corresponding minimized structures [37, 38].

We have also proposed a new type of the main-chain torsion-energy terms for protein systems, which can have amino-acid-dependent force-field parameters [39]. As an example of this formulation, we applied this approach to the AMBER ff03 force field and determined new amino-acid-dependent main-chain torsion-energy parameters for ψ (N-C $_{\alpha}$ -C-N) and ψ' (C $_{\beta}$ -C $_{\alpha}$ -C-N) by using our optimization method in Refs [25, 26, 27].

In this Chapter, we review our works on protein force fields. In section 2 the details of the new main-chain torsion-energy terms and the methods for refinements of force-field parameters are given. In section 3 examples of the applications of these methods are presented. Section 4 is devoted to conclusions.

2 Methods

2.1 General force field for protein systems

The all-atom force fields for protein systems such as AMBER, CHARMM, OPLS, and ECEPP use essentially the same functional forms for the potential energy except for minor differences. The commonly used total conformational potential energy E_{conf} is given by

$$E_{\text{conf}} = E_{\text{BL}} + E_{\text{BA}} + E_{\text{torsion}} + E_{\text{nonbond}} , \quad (1)$$

where

$$E_{\text{BL}} = \sum_{\text{bond length } \ell} K_{\ell} (\ell - \ell_{\text{eq}})^2 , \quad (2)$$

$$E_{\text{BA}} = \sum_{\text{bond angle } \theta} K_{\theta} (\theta - \theta_{\text{eq}})^2 , \quad (3)$$

$$E_{\text{torsion}} = \sum_{\text{dihedral angle } \Phi} \sum_n \frac{V_n}{2} [1 + \cos(n\Phi - \gamma_n)] , \quad (4)$$

$$E_{\text{nonbond}} = \sum_{i < j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{332q_i q_j}{\epsilon r_{ij}} \right] . \quad (5)$$

Here, E_{BL} , E_{BA} , and E_{torsion} represent the bond-stretching term, the bond-bending term, and the torsion-energy term, respectively. The bond-stretching and bond-bending energies are given by harmonic terms with the force constants, K_{ℓ} and K_{θ} , and the equilibrium positions, ℓ_{eq} and θ_{eq} . The torsion energy is, on the other hand, described by the Fourier series in Eq. (5), where the sum is taken over all dihedral

angles Φ , n is the number of waves, γ_n is the phase, and V_n is the Fourier coefficient. The nonbonded energy in Eq. (5) is represented by the Lennard-Jones and Coulomb terms between pairs of atoms, i and j , separated by the distance r_{ij} (in Å). The parameters A_{ij} and B_{ij} in Eq. (5) are the coefficients for the Lennard-Jones term, q_i (in units of electronic charges) is the partial charge of the i -th atom, and ϵ is the dielectric constant, where we usually set $\epsilon = 1$ (the value in vacuum). The factor 332 in the electrostatic term is a constant to express energy in units of kcal/mol. Hence, we have five classes of force-field parameters, namely, those in the bond-stretching term (K_ℓ and ℓ_{eq}), those in the bond-bending term (K_θ and θ_{eq}), those in the torsion term (V_n and γ_n), those in the Lennard-Jones term (A_{ij} and B_{ij}), and those in the electrostatic term (q_i).

Eq. (1) represents a standard set of the potential energy terms. As mentioned above, there are minor differences in the energy functions among different force fields. For instance, the Urey-Bradley term is used in CHARMM and OPLS, but not in AMBER. In our parameter refinement methods, we try to optimize a certain set of parameters in the existing force fields without changing the functional forms. Therefore, if the original force field has non-standard terms, then the optimized one also has them.

2.2 New torsion-energy terms

2.2.1 Representation by a double Fourier series [34, 35]

Separating the contributions $E(\phi, \psi)$ of the backbone dihedral angles ϕ and ψ from the rest of the torsion terms E_{rest} , we can write the torsion energy term in Eq. (5) as

$$E_{torsion} = E(\phi, \psi) + E_{rest} , \quad (6)$$

where we have

$$E(\phi, \psi) = \sum_m \frac{V_m}{2} [1 + \cos(m\phi - \gamma_m)] + \sum_n \frac{V_n}{2} [1 + \cos(n\psi - \gamma_n)] . \quad (7)$$

For example, the coefficients for the cases of six force fields namely, AMBER parm94, AMBER parm96, AMBER parm99, CHARMM27, OPLS-AA, and OPLS-AA/L, are summarized in Table 1, and we can explicitly write $E(\phi, \psi)$ in Eq. (7) as follows:

$$E_{\text{parm94}}(\phi, \psi) = 2.7 - 0.2 \cos 2\phi - 0.75 \cos \psi - 1.35 \cos 2\psi - 0.4 \cos 4\psi \quad (8)$$

$$E_{\text{parm96}}(\phi, \psi) = 2.3 + 0.85 \cos \phi - 0.3 \cos 2\phi + 0.85 \cos \psi - 0.3 \cos 2\psi , \quad (9)$$

$$E_{\text{parm99}}(\phi, \psi) = 5.35 + 0.8 \cos \phi - 0.85 \cos 2\phi - 1.7 \cos \psi - 2.0 \cos 2\psi \quad (10)$$

$$E_{\text{CHARMM}}(\phi, \psi) = 0.8 - 0.2 \cos \phi + 0.6 \cos \psi , \quad (11)$$

$$E_{\text{OPLS-AA}}(\phi, \psi) = 1.158 - 1.1825 \cos \phi - 0.456 \cos 2\phi - 0.425 \cos 3\phi$$

$$+ 0.908 \cos \psi - 0.611 \cos 2\psi + 0.7905 \cos 3\psi, \quad (12)$$

$$E_{\text{OPLS-AA/L}}(\phi, \psi) = 0.81885 - 0.298 \cos \phi - 0.1395 \cos 2\phi - 2.4565 \cos 3\phi \\ + 0.3715 \cos \psi - 1.254 \cos 2\psi - 0.4025 \cos 3\psi. \quad (13)$$

Table 1 Torsion-energy parameters for the backbone dihedral angles ϕ and ψ for AMBER parm94, AMBER parm96, AMBER parm99, CHARMM27, OPLS-AA, and OPLS-AA/L in Eq. (7).

	ϕ			ψ		
	m	$\frac{V_m}{2}$ (kcal/mol)	γ_m (radians)	n	$\frac{V_n}{2}$ (kcal/mol)	γ_n (radians)
parm94	2	0.2	π	1	0.75	π
				2	1.35	π
				4	0.4	π
parm96	1	0.85	0	1	0.85	0
	2	0.3	π	2	0.3	π
parm99	1	0.8	0	1	1.7	π
	2	0.85	π	2	2.0	π
charmm	1	0.2	π	1	0.6	0
opls-aa	1	-1.1825	0	1	0.908	0
	2	0.456	π	2	0.611	π
	3	-0.425	0	3	0.7905	0
opls-aal	1	-0.298	0	1	0.3715	0
	2	0.1395	π	2	1.254	π
	3	-2.4565	0	3	-0.4025	0

The backbone torsion-energy term $E(\phi, \psi)$ in Eq. (7) is a sum of two one-dimensional Fourier series: one is for ϕ and the other for ψ . The two variables ϕ and ψ are decoupled, and no correlation of ϕ and ψ can be incorporated. On the other hand, any periodic function of ϕ and ψ with period 2π can be expanded by a double Fourier series. As a simple generalization of $E(\phi, \psi)$, we therefore proposed to express this backbone torsion energy by the following double Fourier series [34, 35]:

$$\begin{aligned} \mathcal{E}(\phi, \psi) = & a + \sum_{m=1}^{\infty} (b_m \cos m\phi + c_m \sin m\phi) \\ & + \sum_{n=1}^{\infty} (d_n \cos n\psi + e_n \sin n\psi) \\ & + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (f_{mn} \cos m\phi \cos n\psi + g_{mn} \cos m\phi \sin n\psi \\ & + h_{mn} \sin m\phi \cos n\psi + i_{mn} \sin m\phi \sin n\psi). \end{aligned} \quad (14)$$

Here, m and n are the numbers of waves, a , b_m , c_m , d_n , e_n , f_{mn} , g_{mn} , h_{mn} , and i_{mn} are the Fourier coefficients. This equation includes cross terms in ϕ and ψ , while the original term in Eq. (7) has no mixing of ϕ and ψ . Therefore, our new torsion-energy term can represent more complex energy surface than the conventional ones. The Fourier coefficients, by definition, are given by

$$\begin{aligned} c &= \frac{1}{\alpha} \int_{-\pi}^{\pi} d\phi \int_{-\pi}^{\pi} d\psi \mathcal{E}(\phi, \psi) x(\phi, \psi) \\ &= \left(\frac{\pi}{180}\right)^2 \frac{1}{\alpha} \int_{-180}^{180} d\tilde{\phi} \int_{-180}^{180} d\tilde{\psi} \mathcal{E}\left(\frac{\pi}{180}\tilde{\phi}, \frac{\pi}{180}\tilde{\psi}\right) x\left(\frac{\pi}{180}\tilde{\phi}, \frac{\pi}{180}\tilde{\psi}\right), \end{aligned} \quad (15)$$

where α are the normalization constants and $x(\phi, \psi)$ are the basis functions for the Fourier series. Table 2 summarizes these coefficients and functions. Here, ϕ and ψ are given in radians, and $\tilde{\phi}$ and $\tilde{\psi}$ are in degrees ($\phi = \frac{\pi}{180}\tilde{\phi}$, $\psi = \frac{\pi}{180}\tilde{\psi}$). Hereafter, angular quantities without tilde and with tilde are in radians and in degrees, respectively.

Table 2 Fourier coefficients c , normalization constants α , and the basis functions $x(\phi, \psi)$ for the double Fourier series of the backbone torsion energy $\mathcal{E}(\phi, \psi)$ in Eqs. (14) and (15).

c	α	$x(\phi, \psi)$
a	$4\pi^2$	1
b_m	$2\pi^2$	$\cos m\phi$
c_m	$2\pi^2$	$\sin m\phi$
d_n	$2\pi^2$	$\cos n\psi$
e_n	$2\pi^2$	$\sin n\psi$
f_{mn}	π^2	$\cos m\phi \cos n\psi$
g_{mn}	π^2	$\cos m\phi \sin n\psi$
h_{mn}	π^2	$\sin m\phi \cos n\psi$
i_{mn}	π^2	$\sin m\phi \sin n\psi$

Finally, $\mathcal{E}(\phi, \psi)$ in Eq. (14) and E_{rest} in Eq. (6) define our torsion-energy term in Eq. (1) (instead of Eq. (5)):

$$E_{\text{torsion}} = \mathcal{E}(\phi, \psi) + E_{\text{rest}}. \quad (16)$$

The double Fourier series in Eq. (14) is particularly useful, because it describes the backbone torsion-energy surface in the Ramachandran space. The Fourier series can express the torsion-energy surface $\mathcal{E}(\phi, \psi)$ that was obtained by any method including quantum chemistry calculations [16, 28, 29, 30, 13, 14, 31].

Moreover, one can refine the existing backbone torsion-energy term and control the secondary-structure-forming tendencies of the force fields. For example, α -helix is obtained for $(\tilde{\phi}, \tilde{\psi}) \approx (-57^\circ, -47^\circ)$, 3_{10} -helix for $(\tilde{\phi}, \tilde{\psi}) \approx (-49^\circ, -26^\circ)$, π -helix for $(\tilde{\phi}, \tilde{\psi}) \approx (-57^\circ, -70^\circ)$, parallel β -sheet for $(\tilde{\phi}, \tilde{\psi}) \approx (-119^\circ, 113^\circ)$, antiparallel β -sheet for $(\tilde{\phi}, \tilde{\psi}) \approx (-139^\circ, 135^\circ)$, and so on [36]. Hence, if the ex-

isting force field gives, say, too little α -helix-forming tendency compared to experimental results, one can lower the backbone torsion-energy surface near $(\tilde{\phi}, \tilde{\psi}) = (-57^\circ, -47^\circ)$ in order to enhance α -helix formations.

We can thus write

$$\mathcal{E}(\phi, \psi) = E(\phi, \psi) - f(\phi, \psi) , \quad (17)$$

where $E(\phi, \psi)$ is the existing backbone torsion-energy term that we want to refine and $f(\phi, \psi)$ is a function that has peaks around the corresponding regions where specific secondary structures are to be enhanced. There are many possible choices for $f(\phi, \psi)$. For instance, one can use the following function when one wants to lower the torsion-energy surface in a single region near $(\phi, \psi) = (\phi_0, \psi_0)$:

$$f(\phi, \psi) = \begin{cases} A \exp \left(\frac{B}{(\phi - \phi_0)^2 + (\psi - \psi_0)^2 - r_0^2} \right) , & \text{for } (\phi - \phi_0)^2 + (\psi - \psi_0)^2 < r_0^2 , \\ 0 , & \text{otherwise ,} \end{cases} \quad (18)$$

where A , B , and r_0 are constants that we adjust for refinement. In this case, the energy surface is lowered by $f(\phi, \psi)$ in a circular region of radius r_0 , which is centered at $(\phi, \psi) = (\phi_0, \psi_0)$. Note that we should also impose periodic boundary conditions on $f(\phi, \psi)$.

We then express $\mathcal{E}(\phi, \psi)$ in Eq. (17) in terms of the double Fourier series in Eq. (14), where the Fourier coefficients are obtained from Eq. (15). Hence, we can fine-tune the backbone torsion-energy term by the above procedure so that it yields correct secondary-structure-forming tendencies.

Some remark about the computation time is now in order. It may appear that we have to expect great increase in computation time by the introduction of the double Fourier series, because the number of terms are much larger. However, because most of the computation time for the force-field evaluations is spent in the calculations of distances between pairs of atoms in the system, the increase in computation time due to the double Fourier series is essentially negligible compared to these main computational efforts.

2.2.2 Amino-acid-dependent main-chain torsion-energy terms [39]

By writing the dihedral-angle dependence of the parameters explicitly, we can rewrite the torsion-energy term in Eq. (5) as

$$E_{\text{torsion}} = \sum_{\Phi} \sum_n \frac{V_n(\Phi)}{2} \{1 + \cos[n\Phi - \gamma_n(\Phi)]\} , \quad (19)$$

where the first summation is taken over all dihedral angles Φ (both in the main chain and in the side chains), n is the number of waves, γ_n is the phase, and V_n is the Fourier coefficient. Namely, the energy term E_{torsion} has $\gamma_n(\Phi)$ and $V_n(\Phi)$ as force-field parameters.

We can further write the torsion-energy term as

$$E_{\text{torsion}} = E_{\text{torsion}}^{(\text{MC})} + E_{\text{torsion}}^{(\text{SC})}, \quad (20)$$

where $E_{\text{torsion}}^{(\text{MC})}$ and $E_{\text{torsion}}^{(\text{SC})}$ are the torsion-energy terms for dihedral angles around main-chain bonds and around side-chain bonds, respectively. Examples of the dihedral angles in $E_{\text{torsion}}^{(\text{MC})}$ are ϕ (C-N-C $_{\alpha}$ -C), ψ (N-C $_{\alpha}$ -C-N), ϕ' (C $_{\beta}$ -C $_{\alpha}$ -N-C), ψ' (C $_{\beta}$ -C $_{\alpha}$ -C-N), and ω (C $_{\alpha}$ -C-N-C $_{\alpha}$). The force-field parameters in $E_{\text{torsion}}^{(\text{SC})}$ can readily depend on amino-acid residues. However, those in $E_{\text{torsion}}^{(\text{MC})}$ are usually taken to be independent of amino-acid residues and the common parameter values are used for all the amino-acid residues (except for proline). This is because the amino-acid dependence of the force field is believed to be taken care of by the very existence of side chains. In Table 3, we list examples of the parameter values for ψ (N-C $_{\alpha}$ -C-N) and ψ' (C $_{\beta}$ -C $_{\alpha}$ -C-N) in general AMBER force fields.

Table 3 Torsion-energy parameters (V_n and γ_n) for the main-chain dihedral angles ψ and ψ' in Eq. (19) for the original AMBER ff94, ff96, ff99, ff99SB, and ff03 force fields. The values are common among the amino-acid residues for each force field. Only the parameters for non-zero V_n are listed.

force field	ψ (N-C $_{\alpha}$ -C-N)			ψ' (C $_{\beta}$ -C $_{\alpha}$ -C-N)		
	n	$V_n/2$	γ_n	n	$V_n/2$	γ_n
ff94	1	0.75	π	2	0.07	0
	2	1.35	π	4	0.10	0
	4	0.40	π			
ff96	1	0.85	0	2	0.07	0
	2	0.30	π	4	0.10	0
ff99	1	1.70	π	2	0.07	0
	2	2.00	π	4	0.10	0
ff99SB	1	0.45	π	1	0.20	0
	2	1.58	π	2	0.20	0
	3	0.55	π	3	0.40	0
ff03	1	0.6839	π	1	0.7784	π
	2	1.4537	π	2	0.0657	π
	3	0.4615	π	3	0.0560	0

However, this amino-acid independence of the main-chain torsion-energy terms is not an absolute requirement, because we are representing the entire force field by rather a small number of classical-mechanical terms. In order to reproduce the exact quantum-mechanical contributions, one can introduce amino-acid dependence on any force-field term including the main-chain torsion-energy terms. Hence, we can generalize $E_{\text{torsion}}^{(\text{MC})}$ in Eq. (20) from the expression in Eq. (19) to the following amino-acid-dependent form:

$$E_{\text{torsion}}^{(\text{MC})} = \sum_{k=1}^{20} \sum_{\Phi_{\text{MC}}^{(k)}} \sum_n \frac{V_n(\Phi_{\text{MC}}^{(k)})}{2} \left\{ 1 + \cos \left[n\Phi_{\text{MC}}^{(k)} - \gamma_n(\Phi_{\text{MC}}^{(k)}) \right] \right\}, \quad (21)$$

where $k (= 1, 2, \dots, 20)$ is the label for the 20 kinds of amino-acid residues and $\Phi_{\text{MC}}^{(k)}$ are dihedral angles around the main-chain bonds in the k -th amino-acid residue.

2.3 Optimization of force-field parameters

2.3.1 Use of force acting on each atom with the PDB coordinates [25, 26, 27, 40]

In the previous subsection, we presented functional forms of the force fields. Given a fixed set of force-field functions, we try to optimize a certain set of parameters in the force fields without changing the functional forms. Therefore, if the original force field has non-standard terms, then the optimized one also has them.

Our optimization method for these force-field parameters is now described [25]. We first retrieve N native structures (one structure per protein) from PDB. We try to choose proteins from different folds (such as all α -helix, all β -sheet, α/β , etc.) and different homology classes as much as possible. If the force-field parameters are of ideal values, then all the chosen native structures are stable without any force acting on each atom in the molecules on the average. Hence, we expect

$$F = 0, \quad (22)$$

where

$$F = \sum_{m=1}^N \frac{1}{N_m} \sum_{i_m=1}^{N_m} |\mathbf{f}_{i_m}|^2, \quad (23)$$

and

$$\mathbf{f}_{i_m} = -\frac{\partial E_{\text{tot}}^{\{m\}}}{\partial \mathbf{x}_{i_m}}. \quad (24)$$

Here, N_m is the total number of atoms in molecule m , $E_{\text{tot}}^{\{m\}}$ is the total potential energy for molecule m , \mathbf{x}_i is the Cartesian coordinate vector of atom i , and \mathbf{f}_i is the force acting on atom i . In reality, $F \neq 0$, and because $F \geq 0$, we can optimize the force-field parameters by minimizing F with respect to these parameters. In practice, we perform a simulation in the force-field parameter space for this minimization.

Proteins are usually in aqueous solution, and hence we also have to incorporate some kind of solvent effects. Because the more the total number of proteins (N) is, the better the force-field parameter optimizations are expected to be, we want to minimize our efforts in the calculations of the solvent effects. Here, we employ the generalized-Born/surface area (GB/SA) terms for the solvent contributions [41, 42].

Hence, we use in Eq. (24) (we suppress the label m for each molecule)

$$E_{\text{tot}} = E_{\text{conf}} + E_{\text{solv}} , \quad (25)$$

where

$$E_{\text{solv}} = E_{\text{GB}} + E_{\text{SA}} , \quad (26)$$

$$E_{\text{GB}} = -166 \left(1 - \frac{1}{\epsilon_s} \right) \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_{ij}^2} e^{-D_{ij}}} , \quad (27)$$

$$E_{\text{SA}} = \sum_k \sigma_k A_k . \quad (28)$$

Namely, in the GB/SA model, the total solvation free energy in Eq. (26) is given by the sum of a solute-solvent electrostatic polarization term, a solvent-solvent cavity term, and a solute-solvent van der Waals term. A solute-solvent electrostatic polarization term can be calculated by the generalized Born equation in Eq. (27), where $\alpha_{ij} = \sqrt{\alpha_i \alpha_j}$, α_i is the so-called Born radius of atom i , $D_{ij} = r_{ij}^2 / (2\alpha_{ij})^2$, and ϵ_s is the dielectric constant of bulk water (we take $\epsilon_s = 78.3$). A solvent-solvent cavity term and a solute-solvent van der Waals term can be approximated by the term in Eq. (28) that is proportional to the solvent accessible surface area. Here, A_k is the total solvent-accessible surface area of atoms of type k and σ_k is an empirically determined proportionality constant [41, 42].

The flowchart of our method for the optimization of force-field parameters is shown in Fig. 1.

In Step 1 of the flowchart we try to obtain as many structures as possible from PDB. The number is limited by the computer power that we have available in our laboratory. We want to choose proteins with different sizes (numbers of amino acids), different folds, and different homology classes as much as possible. We also want to use only those with high experimental resolutions. Note that only atomic coordinates of proteins are extracted from PDB (and coordinates from other molecules such as crystal water are neglected).

If we use data from X-ray experiments, hydrogen atoms are missing, and thus in Step 2 we have to add hydrogen coordinates. Many protein simulation software packages provide with routines that add hydrogen atoms to the PDB coordinates, and one can use one of such routines.

We now have N protein coordinates ready, but usually such “raw data” result in very high total potential energy and strong forces will be acting on some of the atoms in the molecules. This is because the hydrogen coordinates that we added as above are not based on experimental results and have rather large uncertainties. The coordinates of heavy atoms from PDB also have experimental errors. We take the position that we leave the coordinates of heavy atoms as they are in PDB as much as possible, and adjust the hydrogen coordinates to reduce this mismatch. This is why we want to include as many PDB data as possible with high experimental resolutions (so that the effects of experimental errors in PDB may be minimal). We thus minimize the total potential energy $E_{\text{tot}} = E_{\text{conf}} + E_{\text{solv}} + E_{\text{constr}}$ with respect to

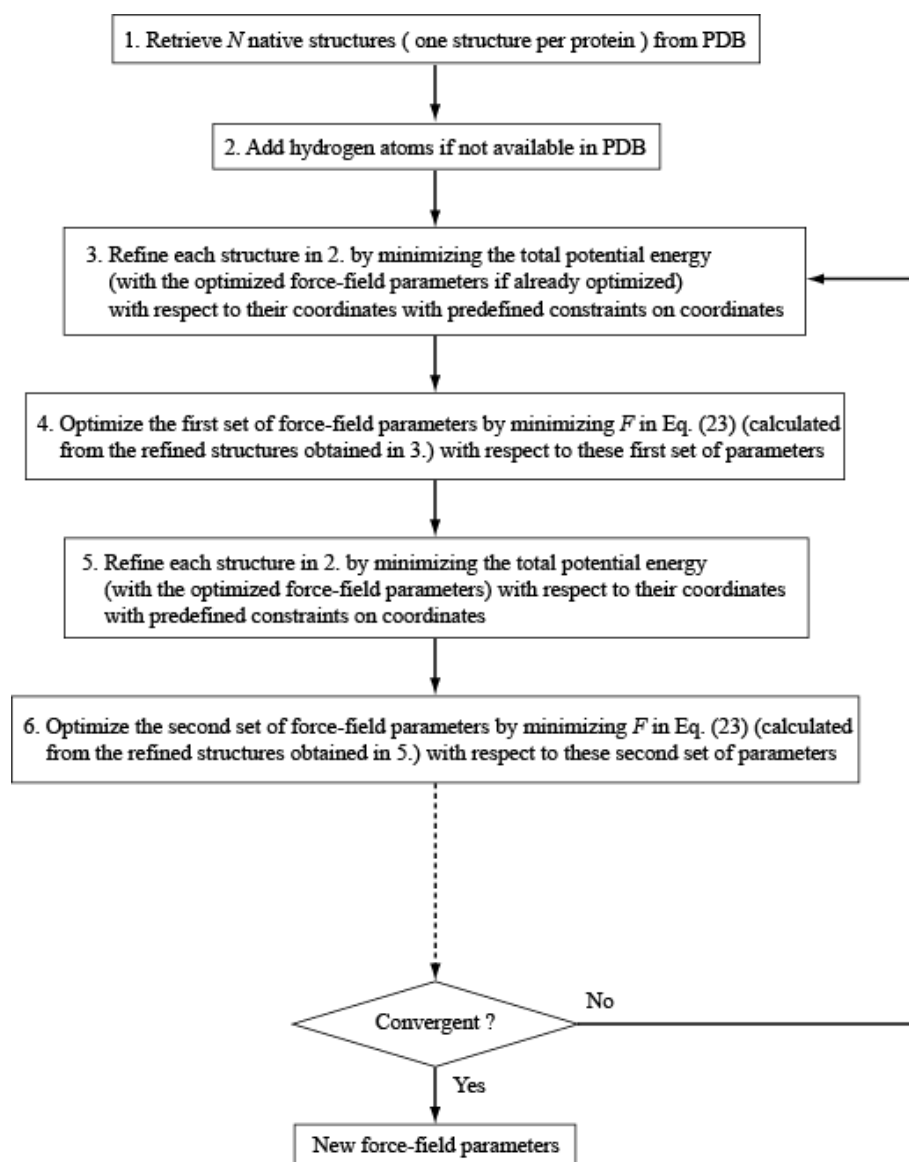


Fig. 1 The flowchart of our method for the optimization of force-field parameters.

the coordinates for each protein conformation, where E_{constr} is the constraint energy term that is imposed on the heavy atoms in PDB (it is referred to as the “predefined constraints” in Steps 3 and 5 in Fig. 1):

$$E_{\text{constr}} = \sum_{\text{heavy atom}} K_x (\mathbf{x} - \mathbf{x}_0)^2. \quad (29)$$

Here, K_x is the force constant of the restriction, and \mathbf{x}_0 are the original coordinate vectors of heavy atoms in PDB. Because we are searching for the nearest local-minimum states, usual minimization routines such as the conjugate-gradient method and Newton-Raphson method can be employed here. As one can see from Eq. (29), the coordinates of hydrogen atoms will be mainly adjusted, but unnatural heavy-atom coordinates will also be modified. We perform this minimization for all N protein structures separately, and obtain N refined structures.

Given N set of “ideal” reference coordinates in Step 3 of the flowchart, we now optimize the first set of force-field parameters in Step 4. In Eq. (1) we have five classes of force-field parameters as mentioned above. Namely, the force-field parameters are those in the bond-stretching term (K_ℓ and ℓ_{eq}), those in the bond-bending term (K_θ and θ_{eq}), those in the torsion term (V_n and γ_n), those in the Lennard-Jones term (A_{ij} and B_{ij}), and those in the electrostatic term (q_i). Because they are of very different nature, we believe that it is better to optimize these classes of force-field parameters separately (as in Steps 4, 6, and so on in Fig. 1). Note also that if we optimize all the parameters simultaneously, the null result (with all the parameter values equal to zero) is a solution to Eq. (22). This is the main reason why we optimize each class of parameters separately.

For each set of force-field parameters, the optimization is carried out by minimizing F in Eq. (23) with respect to these parameters. Here, E_{tot} in Eq. (24) is given by Eq. (25). For this purpose usual minimization routines such as the conjugate-gradient method are not adequate, because we need a global optimization. One should employ more powerful methods such as simulated annealing [43] and generalized-ensemble algorithms [4]. We perform this minimization simulation in the above parameter space to obtain the parameter values that give the global minimum of F .

These processes are repeated until the optimized force-field parameters converge. We can, in principle, optimize all the force-field parameters following the flowchart in Fig. 1. In the examples given below, however, we just optimize two classes of the force-field parameters for simplicity; namely, the partial charges and the backbone torsion-energy parameters. For the optimization of the partial charges (q_i), we impose a condition that the total charge of each amino acid remains constant, which is the usual assumption adopted by the force fields of Eq. (1) based on classical mechanics. As for the main chain torsion-energy parameters, we use the following functional form for each backbone dihedral angle ϕ and ψ (see Eq. (5)):

$$E_{\phi=\phi, \psi} = \frac{V_a}{2} [1 + \cos(n_a \Phi - \gamma_a)] + \frac{V_b}{2} [1 + \cos(n_b \Phi - \gamma_b)] + \frac{V_c}{2} [1 + \cos(n_c \Phi - \gamma_c)]. \quad (30)$$

We optimize only the parameters (V_a , V_b , and V_c) and fix the number of waves (n_a , n_b , and n_c) and the phases (γ_a , γ_b , and γ_c) as in the original force field. This torsion-energy parameter optimization strongly depends on the values of the force constant K_x of the constraint energy in Eq. (29): The larger the values of K_x are, the larger those of V_a , V_b , and V_c tend to be. In order to minimize such dependences, we impose the constraint that the total area enclosed by the curve of $|E_\Phi|$ (from $\Phi = -180^\circ$ to 180°) remains less than or equal to the original value during the optimization.

We believe that these two classes of parameters have the most uncertainty among all the force-field parameters. This is because partial charges are usually obtained by quantum chemistry calculations of an isolated amino acid in vacuum separately, which is a very different condition from that in amino acids of proteins in aqueous solution, and because the torsion-energy term is the most problematic (for instance, the parm94, parm96, and parm99 versions of AMBER differ mainly in backbone torsion-energy parameters).

Moreover, when we perform the optimizations of force-field parameters by using F in Eq. (23), we can neglect unnaturally large forces acting on atoms in order to remove the errors of PDB structures. Namely, we can exclude the term for \mathbf{f}_{im} in Eq. (23) that satisfies

$$|\mathbf{f}_{im}| > f_{\text{cut}}. \quad (31)$$

We determine the cutoff value f_{cut} by using the following function:

$$\Phi\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\Phi_i^{\text{native}} - \Phi_i^{\text{min}})^2}. \quad (32)$$

Here, n is the total number of backbone dihedral angles (ϕ and ψ angles) in all molecules, Φ_i^{native} is the i -th backbone dihedral angle of the native structures and Φ_i^{min} is the corresponding i -th backbone dihedral angle of the minimized structures using the trial force-field parameters. The optimal value of f_{cut} is chosen so that ΦRMSD is the minimal value with $f_{\text{cut}} \leq f_{\text{cut}}^{\text{max}}$, where $f_{\text{cut}}^{\text{max}}$ is obtained in an appropriate way (see an example below).

2.3.2 Use of RMSD I [38]

We now describe another second method for optimizing the force-field parameters. We use N proteins again from PDB, which can be the same proteins as those that we used in the previous optimization method. If the force-field parameters are of ideal values, we expect that all the chosen native structures minimized by the ideal force field do not change after minimizations. Namely, we believe that force-field parameters are better, if they have smaller deviations obtained by minimizations of protein structures. Hence, we expect

$$R = 0, \quad (33)$$

where

$$R = \frac{\sum_{i=1}^N RMSD_i}{N}. \quad (34)$$

Here, $RMSD_i$ is the root-mean-square deviation between the native structure of protein i and the corresponding minimized structure using the trial force-field parameters. In reality, $R \neq 0$, and because $R \geq 0$, we expect that we can optimize the force-field parameters by minimizing R with respect to these force-field parameters. In practice, we perform a simulation in the force-field parameter space for this minimization. Namely, in the previous method we minimize F in Eq. (23), and in the present method we minimize R in Eq. (34) instead.

2.3.3 Use of RMSD II [37]

We now describe our third method for optimizing the force-field parameters. We first select N proteins from PDB as in the previous two methods. If the force-field parameters are of ideal values, we expect that all the chosen native structures minimized by the ideal force field do not change. Namely, we believe that force-field parameters are better, if they have lower deviations obtained from minimizations of protein structures. Hence, we expect

$$\Phi RMSD = 0, \quad (35)$$

where

$$\Phi RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n (\Phi_i^{\text{native}} - \Phi_i^{\text{min}})^2}. \quad (36)$$

Here, n is the total number of backbone dihedral angles (ϕ and ψ angles) in all molecules, Φ_i^{native} is the i -th backbone dihedral angle of the native structures and Φ_i^{min} is the corresponding i -th backbone dihedral angle of the minimized structures using the trial force-field parameters. In reality, $\Phi RMSD \neq 0$, because $\Phi RMSD \geq 0$, we expect that we can optimize the force-field parameters by minimizing $\Phi RMSD$ with respect to these force-field parameters. In practice, we perform a simulation in the force-field parameter space for this minimization.

However, our first aim is to determine the balance of secondary-structure-forming tendencies such as helix structure and β -sheet structure. Additionally, it is difficult to perform the minimization of $\Phi RMSD$ in wider force-field parameter space until $\Phi RMSD$ is close to 0 because of the computational cost. Therefore, we only focus on secondary-structure regions of helix structure and β -sheet structure in the amino-acid sequence. Namely, we only consider the backbone dihedral angles of residues in the native structures which are identified by the DSSP program [44] that they constitute one of α -helix, 3/10-helix, π -helix, and β -sheet structures. We calculate two kinds of $\Phi RMSD$ for secondary structures, namely, $\Phi RMSD_{\text{helix}}$ and $\Phi RMSD_{\beta}$. Here, $\Phi RMSD_{\text{helix}}$ stands for $\Phi RMSD$ of backbone dihedral angles of residues which have helix structures in the native structures, and $\Phi RMSD_{\beta}$ means

that of only β -sheet structures in the native structures. Using these two Φ RMSDs, we want to optimize the torsion-energy parameters, which will have better balance of secondary-structure-forming tendencies. We propose the following combination:

$$\Phi\text{RMSD}_{2\text{ndly}} = \lambda \Phi\text{RMSD}_{\text{helix}} + \Phi\text{RMSD}_{\beta}, \quad (37)$$

where we have introduced a fixed scaling factor λ .

Finally, by minimizing $\Phi\text{RMSD}_{2\text{ndly}}$ with respect to the force-field parameters, we can obtain the optimized force-field parameters.

2.3.4 Use of short MD simulations [45]

We now describe our fourth method for optimizing the force-field parameters. In this method, we prepare M protein structures, which are some experimentally determined conformations. For these proteins, we perform MD simulations, which start from the experimental conformations, by using a trial force field. We try to perform MD simulations with varied values of force-field parameters. After that, we estimate the “ S ” value defined by the following function for the trajectories of the M proteins obtained from the trial MD simulations:

$$S = \sum_{i=1}^M \left(\frac{n_i^{\text{S} \rightarrow \text{U}}}{N_i^{\text{S}}} + \frac{n_i^{\text{U} \rightarrow \text{S}}}{N_i^{\text{U}}} \right). \quad (38)$$

Here, $n_i^{\text{S} \rightarrow \text{U}}$ is the number of the amino acids in protein i where their structures in PDB (initial conformation) had some secondary structures (such as α -helix, 3_{10} -helix, π -helix, and β structures) but transformed into unstructured, coil structures without any secondary structures after a short MD simulation. Likewise, $n_i^{\text{U} \rightarrow \text{S}}$ is the number of amino acids in protein i where their structures in PDB had coil structures but transformed to have some secondary structures after a MD simulation. N_i^{S} is the total number of amino acids in protein i which have some secondary structures in PDB, and N_i^{U} is the total number of amino acids in protein i which have coil structures in PDB.

When we calculate the S values for the conformations obtained from MD simulations by using trial force-field parameters, the parameter set, which yields the minimum S value, is considered to give the optimized force field.

3 Examples of Optimizations of Force-Field Parameters

3.1 New torsion-energy terms

3.1.1 Representation by a double Fourier series [34, 35]

We now present various examples of our refinements of force-field parameters. We first consider the following truncated double Fourier series (see Eq. (14)):

$$\begin{aligned}
 \mathcal{E}(\phi, \psi) = & a + b_1 \cos \phi + c_1 \sin \phi + b_2 \cos 2\phi + c_2 \sin 2\phi + b_3 \cos 3\phi + c_3 \sin 3\phi \\
 & + d_1 \cos \psi + e_1 \sin \psi + d_2 \cos 2\psi + e_2 \sin 2\psi + d_3 \cos 3\psi + e_3 \sin 3\psi \\
 & + f_{11} \cos \phi \cos \psi + g_{11} \cos \phi \sin \psi + h_{11} \sin \phi \cos \psi + i_{11} \sin \phi \sin \psi \\
 & + f_{21} \cos 2\phi \cos \psi + g_{21} \cos 2\phi \sin \psi + h_{21} \sin 2\phi \cos \psi + i_{21} \sin 2\phi \sin \psi \\
 & + f_{12} \cos \phi \cos 2\psi + g_{12} \cos \phi \sin 2\psi + h_{12} \sin \phi \cos 2\psi + i_{12} \sin \phi \sin 2\psi \\
 & + f_{22} \cos 2\phi \cos 2\psi + g_{22} \cos 2\phi \sin 2\psi \\
 & + h_{22} \sin 2\phi \cos 2\psi + i_{22} \sin 2\phi \sin 2\psi .
 \end{aligned} \tag{39}$$

This function has 29 Fourier-coefficient parameters. We will see below that this number of Fourier terms is sufficient for most of our purposes.

We first check how well the truncated Fourier series in Eq. (39) can reproduce the six original backbone torsion-energy terms in Eqs. (8)–(13). Because these functions are already the sum of one-dimensional Fourier series and subsets of the double Fourier series in Eq. (14), the Fourier coefficients in Eq. (15) can be analytically calculated and agree with those in Eqs. (8)–(13) except for the last one (that for $\cos 4\psi$) in Eq. (8). This term is missing in Eq. (39). These cases thus give us good test of numerical integrations in Eq. (15). The numerical integrations were evaluated as follows. We divided the Ramachandran space ($-180^\circ < \tilde{\phi} < 180^\circ$, $-180^\circ < \tilde{\psi} < 180^\circ$) into unit square cells of side length $\tilde{\epsilon}$ (in degrees). Hence, there are $(360/\tilde{\epsilon})^2$ unit cells altogether. The double integral on the right-hand side of Eq. (15) was approximated by the sum of $[\mathcal{E}(\frac{\pi}{180}\tilde{\phi}, \frac{\pi}{180}\tilde{\psi}) \times (\frac{\pi}{180}\tilde{\phi}, \frac{\pi}{180}\tilde{\psi})] \times (\tilde{\epsilon})^2$, where each $\mathcal{E}(\frac{\pi}{180}\tilde{\phi}, \frac{\pi}{180}\tilde{\psi})$ was evaluated at one of the four corners of each unit cell. We tried two values of $\tilde{\epsilon}$ (1° and 10°). Both cases gave almost complete agreement of Fourier coefficients with the results of the analytical integrations (see, for example, Tables 4 below).

In Fig. 2 we compare the six original backbone torsion-energy surfaces with those of the corresponding double Fourier series in Eq. (39). Hereafter, the primed labels for figures such as (a') indicate that the results are those of the double Fourier series. As can be seen from Fig. 2, the backbone torsion-energy surfaces are in complete agreement for all force fields except for AMBER parm94, whereas we see a little difference for AMBER parm94 between Figs. 2(a) and 2(a'). As discussed above, this slight difference for AMBER parm94 reflects the fact that the $\cos 4\psi$ term in Eq. (8) is missing in the truncated double Fourier series in Eq. (39).

Table 4 Fourier coefficients in Eq. (39) obtained from the numerical evaluations of the integrals in Eq. (15). “org94” stands for the original AMBER parm94 force field. “mod94(α)” and “mod94(β)” stand for AMBER parm94 force fields that were modified to enhance α -helix structures and β -sheet structures, respectively, by Eqs. (17) and (18). The bin size $\tilde{\epsilon}$ is the length of the sides of each unit square cell for the numerical integration in Eq. (15).

bin size $\tilde{\epsilon}$	1°			10°		
coefficient	org94	mod94(α)	mod94(β)	org94	mod94(α)	mod94(β)
a	2.700000	2.308359	1.916719	2.700000	2.308370	1.916742
b_1	0.000000	-0.330937	0.781150	0.000000	-0.331053	0.781041
c_1	0.000000	0.509599	0.930938	0.000000	0.509517	0.930809
b_2	-0.200000	-0.101549	-0.115937	-0.200000	-0.101513	-0.115970
c_2	0.000000	0.221123	-0.476745	0.000000	0.221100	-0.476558
b_3	0.000000	-0.018073	0.031693	0.000000	-0.018084	0.031714
c_3	0.000000	-0.002862	-0.018298	0.000000	-0.003036	-0.018310
d_1	-0.750000	-1.164401	-0.052959	-0.750000	-1.164500	-0.052874
e_1	0.000000	0.444390	-0.995478	0.000000	0.444289	-0.995599
d_2	-1.350000	-1.333115	-1.184428	-1.350000	-1.333073	-1.184340
e_2	0.000000	0.241460	0.454905	0.000000	0.241451	0.455147
d_3	0.000000	-0.014220	0.035349	0.000000	-0.014143	0.035324
e_3	0.000000	-0.011515	0.009472	0.000000	-0.011671	0.009465
f_{11}	0.000000	-0.342789	-0.680493	0.000000	-0.343087	-0.680497
g_{11}	0.000000	0.367596	0.971845	0.000000	0.367697	0.971851
h_{11}	0.000000	0.527849	-0.810980	0.000000	0.527949	-0.810985
i_{11}	0.000000	-0.566049	1.158199	0.000000	-0.565751	1.158206
f_{21}	0.000000	0.090016	-0.064642	0.000000	0.090168	-0.064636
g_{21}	0.000000	-0.096530	0.092318	0.000000	-0.096472	0.092309
h_{21}	0.000000	0.202178	0.366601	0.000000	0.202421	0.366565
i_{21}	0.000000	-0.216810	-0.523561	0.000000	-0.216596	-0.523509
f_{12}	0.000000	0.012329	-0.142682	0.000000	0.012385	-0.142712
g_{12}	0.000000	0.176308	-0.392017	0.000000	0.176622	-0.392098
h_{12}	0.000000	-0.018984	-0.170042	0.000000	-0.019013	-0.170077
i_{12}	0.000000	-0.271490	-0.467187	0.000000	-0.271321	-0.467284
f_{22}	0.000000	-0.000586	-0.002453	-0.000001	-0.000585	-0.002451
g_{22}	0.000000	-0.008378	-0.006738	0.000000	-0.008397	-0.006733
h_{22}	0.000000	-0.001316	0.013909	0.000000	-0.001317	0.013897
i_{22}	0.000000	-0.018817	0.038215	0.000000	-0.018867	0.038183

We now consider the double Fourier series of non-trigonometric functions. The functions are those in Eqs. (17) and (18). We try to fine-tune the six original force fields by subtracting $f(\phi, \psi)$ in Eq. (18) from the original functions. The criterion for fine-tuning is, for instance, whether the refined force fields yield better agreement of the secondary-structure-forming tendencies with experimental implications than the original ones. For this we need good experimental data. Because the purpose here is to test whether or not we can control the secondary-structure-forming tendencies, we simply consider extreme cases where we try to modify the existing force fields so that desired secondary structures may be obtained regardless of the

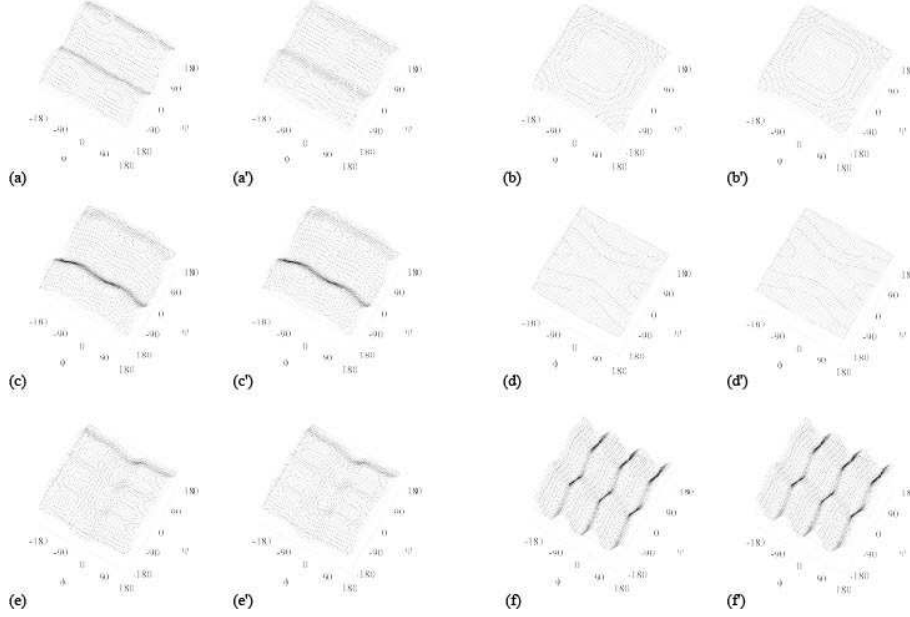


Fig. 2 Backbone-torsion-energy surfaces of six force fields. The backbone dihedral angles $\tilde{\phi}$ and $\tilde{\psi}$ are in degrees. (a), (b), (c), (d), (e), and (f) are those of the original AMBER parm94, the original AMBER parm99, the original CHARMM 27, the original OPLS-AA, and the original OPLS-AA/L, respectively. (a') to (f') are those of (a) to (f), respectively, that were expressed by the truncated double Fourier series in Eq. (39). The contour lines are drawn every 0.5 kcal/mol.

tendencies of the original force fields. Note that the six original force fields have quite different preferences for α -helix and β -sheet structures [23, 24, 25, 26, 27].

The function $f(\phi, \psi)$ in Eq. (18) reduces the value of $E(\phi, \psi)$ in a circle of radius r_0 with the center located at (ϕ_0, ψ_0) . We used $\tilde{r}_0 = 100^\circ$ and $\tilde{B} = 5,000$ (degrees)². The coefficient A is calculated by Eq. (18) from the other parameters $f(\tilde{\phi}_0, \tilde{\psi}_0)$, \tilde{r}_0 , and \tilde{B} . Namely, we have

$$A = f(\tilde{\phi}_0, \tilde{\psi}_0) \exp\left(\frac{\tilde{B}}{\tilde{r}_0^2}\right). \quad (40)$$

We used $(\tilde{\phi}_0, \tilde{\psi}_0) = (-57^\circ, -47^\circ)$ and $(\tilde{\phi}_0, \tilde{\psi}_0) = (-130^\circ, 125^\circ)$ in order to enhance α -helix-forming tendency and β -sheet-forming tendency, respectively. The central values $f(\tilde{\phi}_0, \tilde{\psi}_0)$ that we used were 3.0 kcal/mol and 6.0 kcal/mol for enhancing α -helix and β -sheet, respectively, in the case of AMBER parm94, AMBER parm99, CHARMM27, and OPLS-AA/L. They were both 3.0 kcal/mol in the case of AMBER parm96 and OPLS-AA.

We remark that the large value of $f(\tilde{\phi}_0, \tilde{\psi}_0)$, 6.0 kcal/mol, that was necessary to enhance β -sheet in the case of AMBER parm94, AMBER parm99, CHARMM27, and OPLS-AA/L reflects the fact that their original force fields favor α -helix.

In Fig. 3(a1)–(f1) we compare the six backbone torsion-energy surfaces modified according to Eq. (17), which reduced the torsion energy in the α -helix region, with those of the corresponding double Fourier series in Eq. (39). In Fig. 3(a1)–(f1), α -helix is enhanced from the original AMBER parm94 (a1), AMBER parm96 (b1), AMBER parm99 (c1), CHARMM27 (d1), OPLS-AA (e1), and OPLS-AA/L (f1). In Fig. 4(a1)–(f1) we show the case of the β -sheet region, and β -sheet is enhanced from the original AMBER parm94 (a1), AMBER parm96 (b1), AMBER parm99 (c1), CHARMM27 (d1), OPLS-AA (e1), and OPLS-AA/L (f1).

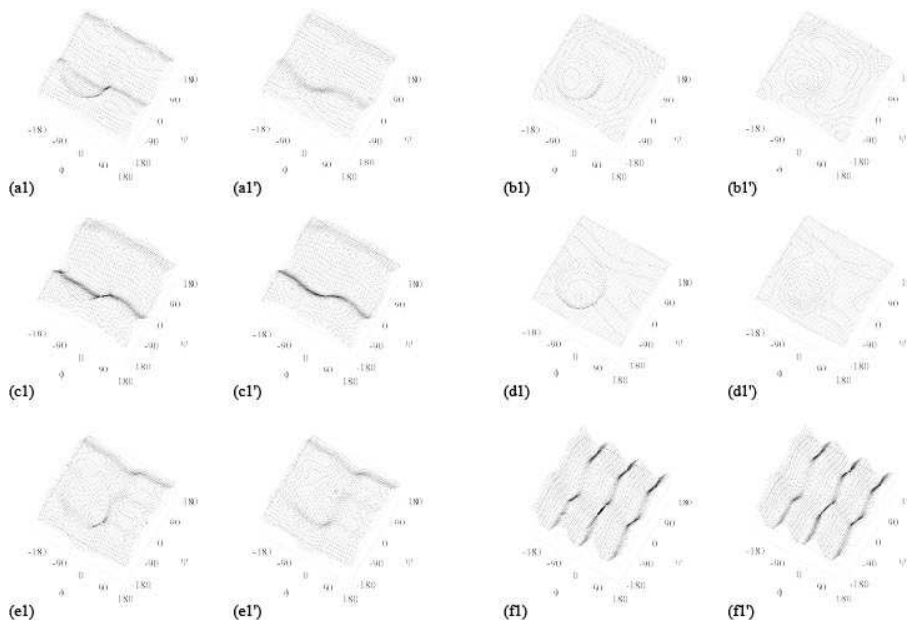


Fig. 3 Backbone-torsion-energy surfaces of six force fields that were modified by Eqs. (17), (18 and (39)). From (a1) to (f1) are those of AMBER parm94, AMBER parm96, AMBER parm99, CHARMM 27, OPLS-AA, and OPLS-AA/L force fields that were modified to enhance α -helix structures, respectively. From (a1') to (f1') are those of AMBER parm94, AMBER parm96, AMBER parm99, CHARMM 27, OPLS-AA, and OPLS-AA/L force fields that were expanded by the truncated double Fourier series in Eq. (39).

These modified backbone torsion-energy functions were expanded by the truncated double Fourier series in Eq. (39) by evaluating the corresponding Fourier coefficients from Eq. (15). For the numerical integration we again tried two values of the bin size $\tilde{\epsilon}$ (1° and 10°). The obtained Fourier coefficients are summarized in Tables 4, for example, in the case of AMBER parm94. For comparisons, the

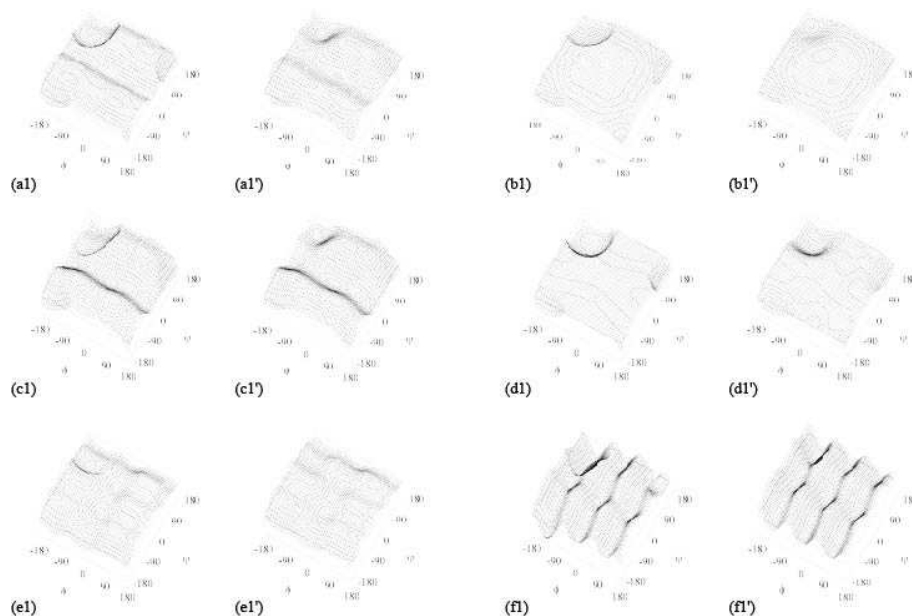


Fig. 4 Backbone-torsion-energy surfaces of six force fields that were modified by Eqs. (17), (18 and (39)). From (a1) to (f1) are those of AMBER parm94, AMBER parm96, AMBER parm99, CHARMM 27, OPLS-AA, and OPLS-AA/L force fields that were modified to enhance β -sheet structures, respectively. From (a1') to (f1') are those of AMBER parm94, AMBER parm96, AMBER parm99, CHARMM 27, OPLS-AA, and OPLS-AA/L force fields that were expanded by the truncated double Fourier series in Eq. (39).

Fourier coefficients of the original AMBER force fields (before modifications) are also listed. We see that the two choices of the bin size $\tilde{\epsilon}$ gave essentially the same results (agreeing in about 3 digits).

In Figs. 3(a1')–(f1') and 4(a1')–(f1') we show the backbone torsion-energy surfaces represented by the truncated double Fourier series. Comparing these with the original ones in Fig. 3(a1)–(f1) and 4(a1)–(f1), we find that the overall features of the energy surfaces are well reproduced by the Fourier series. If more accuracy is desired, we can simply increase the number of Fourier terms in the expansion. As we will see below, the present accuracy of the Fourier series was sufficient for the purpose of controlling the secondary-structure-forming tendencies towards α -helix or β -sheet.

We examined the effects of the above modifications of the backbone torsion-energy terms in AMBER parm94, AMBER parm96, AMBER parm99, CHARMM27, OPLS-AA, and OPLS-AA/L (towards specific secondary structures) by performing the folding simulations of two peptides, namely, C-peptide of ribonuclease A and the C-terminal fragment of the B1 domain of streptococcal protein G, which is sometimes referred to as G-peptide [46]. The C-peptide has 13 residues and its amino-acid sequence is Lys-Glu-Thr-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-Met.

This peptide has been extensively studied by experiments and is known to form an α -helix structure [47, 48], as shown in Fig. 5(a). Because the charges at peptide termini are known to affect helix stability [47, 48], we blocked the termini by a neutral COCH_3 - group and a neutral $-\text{NH}_2$ group. The G-peptide has 16 residues and its amino-acid sequence is Gly-Glu-Trp-Thr-Tyr-Asp-Asp-Ala-Thr-Lys-Thr-Phe-Thr-Val-Thr-Glu. The termini were kept as the usual zwitter ionic states, following the experimental conditions [46, 49, 50]. This peptide is known to form a β -hairpin structure by experiments [46, 49, 50], as shown in Fig. 5(b).

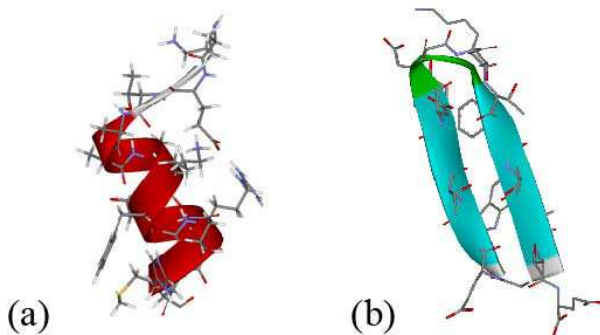


Fig. 5 The structures of C-peptide (a) and G-peptide (b) obtained from the experimental results (PDB ID are (a) 1A5P and (b) 1PGA). The figures were created with DS Visualizer v1.5[51].

Simulated annealing [43] MD simulations were performed for both peptides from fully extended initial conformations, where the 12 versions of the truncated double Fourier series (which were described in Table 4 and in Fig. 3(a1')–(f1') and 4(a1')–(f1')) were used for the backbone torsion-energy terms of AMBER parm94, AMBER parm96, AMBER parm99, CHARMM27, OPLS-AA, and OPLS-AA/L force fields. For comparisons, the simulations with the original force fields were also performed. The unit time step was set to 1.0 fs. Each simulation was carried out for 1 ns (hence, it consisted of 1,000,000 MD steps). The temperature during MD simulations was controlled by Berendsen's method [52]. For each run the temperature was decreased exponentially from 2,000 K to 250 K. We modified and used the program package TINKER version 4.1 [53] for all the simulations. As for solvent effects, we used the GB/SA model [41, 42] included in the TINKER program package. For both peptides, these folding simulations were repeated 60 times with different sets of randomly generated initial velocities.

In Fig. 6, we show seven (out of 60) lowest-energy final conformations of C-peptide and G-peptide obtained by the simulated annealing MD simulations, for example, in the case of AMBER parm94.

In the Figure, we see that all conformations of the original AMBER parm94 (except for conformations 2 and 4 of G-peptide) and all conformations of its force field modified towards α -helix are α -helix structures (conformations 2 and 4 are

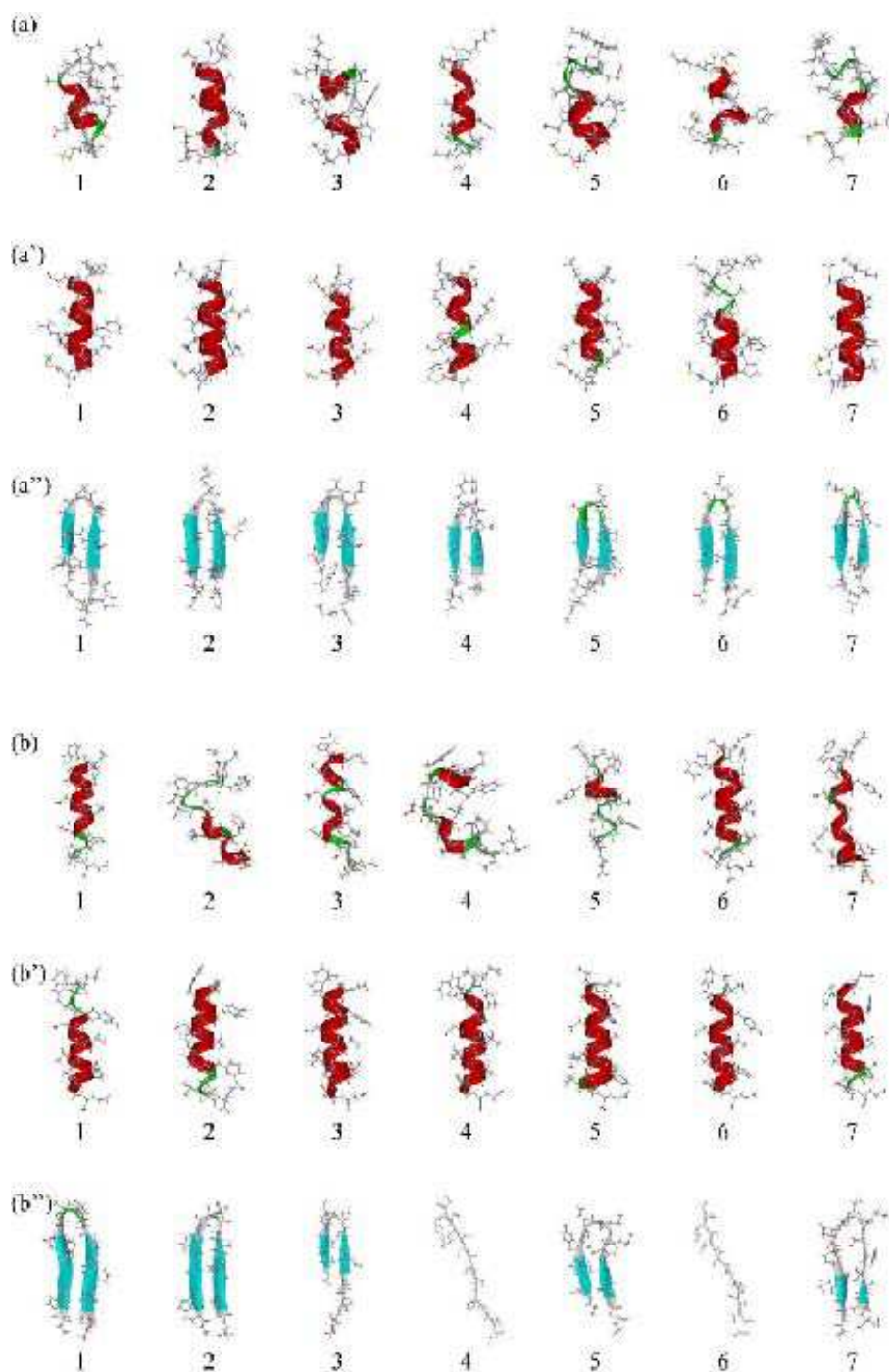


Fig. 6 Seven lowest-energy final conformations of C-peptide (a)–(a'') and G-peptide (b)–(b'') obtained from six sets of 60 simulated annealing MD runs. (a) and (b) are the results of the original AMBER parm94. (a') and (b') are the results of AMBER parm94 of the truncated double Fourier series of six force fields that were modified to enhance α -helix structures. (a'') and (b'') are the results of AMBER parm94 of the truncated double Fourier series of six force fields that were modified to enhance β -sheet structures. The conformations are ordered in the increasing order of energy for each case. The figures were created with DS Visualizer v1.5[54].

3_{10} -helix structures). The results show that the original AMBER parm94 favors α -helix structures, and moreover, its force field modified towards α -helix favors α -helix structures more than the original force field in the sense that the obtained helices are more extended (and almost entirely helical). On the other hand, AMBER parm94 modified towards β -sheet favors β structures strongly. The results for other force fields were similar.

Therefore, regardless of the secondary-structure-forming tendencies of the original force fields, our modifications of the backbone torsion-energy term succeeded in enhancing the desired secondary structures.

3.1.2 Amino-acid-dependent main-chain torsion-energy terms [39]

We present the results of our optimizations of the force-field parameters $V_1(\Phi_{MC}^{(k)})$ for the main-chain angles $\Phi_{MC}^{(k)} = \psi^{(k)}$ (N-C $_{\alpha}$ -C-N) and $\psi'^{(k)}$ (C $_{\beta}$ -C $_{\alpha}$ -C-N) in Eq. (21). We did this for the case of AMBER ff03 force field. We determined these $V_1(\Phi_{MC}^{(k)})$ values for the 19 amino-acid residues except for proline.

At first, we chose 100 PDB files with resolution 2.0 Å or better, with sequence similarity of amino acid 30.0 % or lower, and with less than 200 residues (the average number of residues is 117.0) from PDB-REPRDB [55] (see Table 5 and Fig. 7). We then refined these selected 100 structures. Generally, data from X-ray experiments do not have coordinates for hydrogen atoms. Therefore, we have to add hydrogen coordinates. Many protein simulation software packages provide with routines that add hydrogen atoms to the PDB coordinates. We used the AMBER11 program package [56]. We thus minimized the total potential energy $E_{\text{total}} = E_{\text{conf}} + E_{\text{solv}} + E_{\text{constr}}$ with respect to the coordinates for each protein conformation, where E_{constr} is the harmonic constraint energy term ($E_{\text{constr}} = \sum_{\text{heavy atom}} K_x (\mathbf{x} - \mathbf{x}_0)^2$), and E_{solv} is the solvation energy term. Here, K_x is the force constant of the restriction and \mathbf{x}_0 are the original coordinate vectors of heavy atoms in PDB. As one can see from E_{constr} , the coordinates of hydrogen atoms will be mainly adjusted, but unnatural heavy-atom coordinates will also be modified. We performed this minimization for all the 100 protein structures separately and obtained 100 refined structures by using $K_x = 100$ (kcal/mol). As for the solvation energy term E_{solv} , we used the GB/SA solvent included in the AMBER program package ($igb = 5$ and $gbsa = 1$) [57, 58].

For these refined protein structures, we performed the optimization of force-field parameters $V_1^{(k)}$ of ψ and ψ' angles for AMBER ff03 force field by using the function F in Eq. (23) as the total potential energy function ($E_{\text{total}} = E_{\text{conf}} + E_{\text{solv}}$) for the Monte Carlo simulations in the parameter space. Here, we used AMBER11 [56] for the force calculations in Eq. (24). We have to optimize the 38 ($= 2 \times 19$) parameters simultaneously by the simulations in 38 parameters. However, here, for simplicity, we just optimized two parameters, $V_1(\psi^{(k)})$ and $V_1(\psi'^{(k)})$, for each amino-acid residue k separately, keeping the other V_1 values as the original values. In order to obtain the optimal parameters, we performed Monte Carlo simulations of two pa-

Table 5 100 proteins used in the optimization of force-field parameters.

fold	PDB ID	chain	PDB ID	chain	PDB ID	chain	PDB ID	chain
all α	1DLW	A	1N1J	B	1U84	A	1HBK	A
	1TX4	A	1V54	E	1SK7	A	1TQG	A
	1V74	B	1DVO	A	1HFE	S	1J0P	A
	1Y02	A71-114	1IJY	A	1I2T	A	1G8E	A
	1VKE	C	1FS1	A109-149	1D9C	A	1AIL	A
	1Q5Z	A	1T8K	A	1OR7	C	1NG6	A
	1C75	A	2LIS	A	1NH2	B	1Q2H	A
	1NKP	A						
all β	1XAK	A	1T2W	A	1GMU	C1-70	1AYO	A
	1PK6	A	1NLQ	C	1BEH	A	1UA8	A
	1UXZ	A	1UB4	C	1LGP	A	1CQY	A
	1PM4	A	1OU8	A	1V76	A	1UT7	B
	1OA8	D	1IFG	A				
α/β	1IO0	A	1U7P	A	1JKE	C	1MXI	A
	1LY1	A	1NRZ	A	1IM5	A	1VC1	A
	1OGD	A	1IIB	A	1PYO	D	1MUG	A
	1H75	A	1K66	A	1COZ	A	1D4O	A
$\alpha + \beta$	1VCC	A	1PP0	B	1PZ4	A	1TU1	A
	1Q2Y	A	1M4J	A	1N9L	A	1LQV	B
	1A3A	A	1K2E	A	1TT8	A	1HUF	A
	1SXR	A	1CYO	A	1KAF	A	1ID0	A
	1UCD	A	1F46	B	1KPF	A	1BYR	A
	1Y60	D	1SEI	A	1RL6	A	1WM3	A
	1FTH	A	1APY	B	1JID	A	1N13	E
	1LTS	C	1JYO	F	1E87	A	1UGI	A
	1MWP	A	1PCF	A	1MBY	A	1IHR	B
	1H6H	A						

rameters (V_1 of ψ and ψ') for the 19 amino-acid residues except for proline. In Table 6, the optimized parameters are listed.

In order to check the force-field parameters obtained by our optimization method, we performed the folding simulations using two peptides, namely, C-peptide and G-peptide.

For the folding simulations, we used replica-exchange molecular dynamics (REMD) [59]. REMD is one of the generalized-ensemble algorithms, and has high conformational sampling efficiency by allowing configurations to heat up and cool down while maintaining proper Boltzmann distributions. We used the AMBER11 program package [56]. The unit time step was set to 2.0 fs, and the bonds involving hydrogen atoms were constrained by SHAKE algorithm [60]. Each simulation was carried out for 30.0 ns (hence, it consisted of 15,000,000 MD steps) with 16 replicas by using Langevin dynamics. The exchange procedure for each replica were performed every 3,000 MD steps. The temperature was distributed exponentially: 650, 612, 577, 544, 512, 483, 455, 428, 404, 380, 358, 338, 318, 300, 282, and 266 K. As for solvent effects, we used the GB/SA model in the AMBER program package ($igb = 5$ and $gbsa = 1$) [57, 58]. The initial conformations for each peptide were fully extended

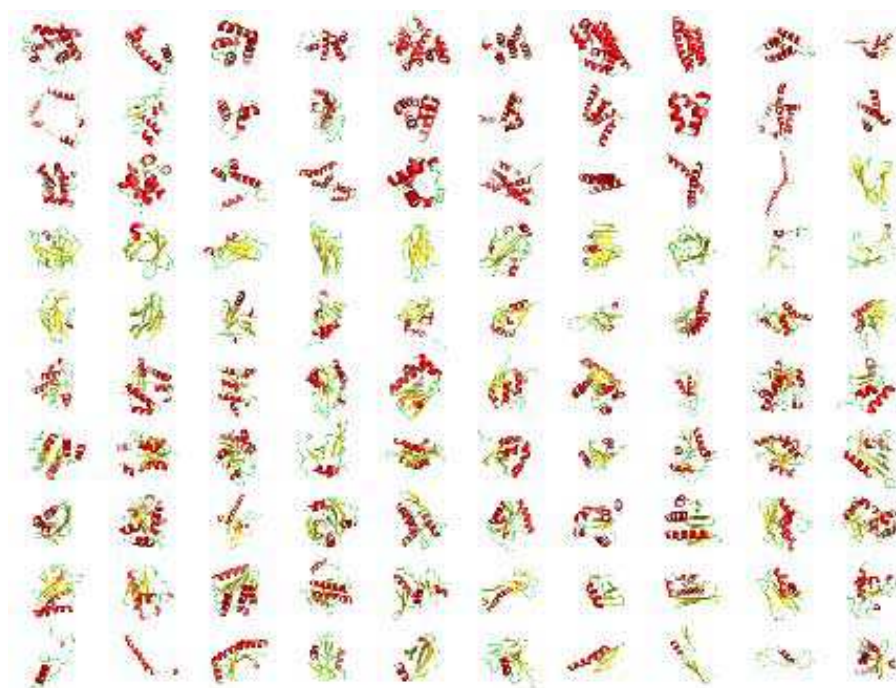


Fig. 7 Structures of 100 proteins in Table 5 which were used in the optimization of force-field parameters.

ones for all the replicas. The REMD simulations were performed with different sets of randomly generated initial velocities for each replica.

In Fig. 8, α -helicity and β -strandness of the two peptides obtained from the REMD simulations are shown. We checked the secondary-structure formations by using the DSSP program [44], which is based on the formations of the intra-main-chain hydrogen bonds. As is shown in Fig. 8, for the original AMBER ff03 force field, the α -helicity is clearly higher than the β -strandness not only in C-peptide but also in G-peptide. Namely, the original AMBER ff03 force field clearly favors α -helix and does not favor β -structure. On the other hand, for the optimized force field, in the case of C-peptide, the α -helicity is higher than the β -strandness, and in the case of G-peptide, the β -strandness is higher than the α -helicity. We conclude that these results obtained from the optimized force field are in better agreement with the experimental results in comparison with the original force field. In Fig. 9, 3_{10} -helicity and π -helicity of two peptides obtained from the REMD simulations are shown. For 3_{10} helicity, there is no large difference for both force fields in C-peptide, and in the case of G-peptide, the value of the optimized force field slightly decreases in comparison with the original force field. π -helicity has almost no value in the both cases of the original and optimized force fields in two peptides.

Table 6 Optimized $V_1/2$ parameters for the main-chain dihedral angles ψ and ψ' for the 19 amino-acid residues (except for proline) in Eq. (21). The rest of the parameters are taken to be the same as in the original ff03 force field. The original amino-acid-independent values are also listed for reference.

	ψ (N-C α -C-N)	ψ' (C β -C α -C-N)
original ff03	0.6839	0.7784
Ala	0.122	0.150
Arg	0.409	0.200
Asn	-0.074	-0.162
Asp	-0.137	0.182
Cys	0.361	0.089
Gln	0.144	-0.024
Glu	0.180	0.152
Gly	0.258	—
His	0.020	0.237
Ile	0.643	0.194
Leu	0.382	0.257
Lys	0.222	0.042
Met	0.141	0.346
Phe	-0.010	0.553
Ser	-0.248	0.475
Thr	0.512	0.328
Trp	0.027	0.477
Tyr	0.082	0.652
Val	0.142	0.590

In Fig. 10, α -helicity and β -strandness as functions of temperature for the two peptides obtained from the REMD simulations are shown. For α -helicity, the values of both force fields decrease gradually from low temperature to high temperature in the case of C-peptide. On the other hand, in the case of G-peptide, there are small peaks at around 300 K and 358 K for the original and optimized force fields, respectively. For β -strandness, in the case of C-peptide, it is almost zero for both force fields. In the case of G-peptide, for the optimized force field, there is clearly a peak around 300 K.

3.2 Optimization of force-field parameters

3.2.1 Use of force acting on each atom in the PDB coordinates [25, 26, 27, 40]

We now present the results of our force-field optimizations. In Step 1 of the flowchart in Fig. 1, we chose 100 PDB files ($N = 100$) from X-ray experiments with resolution 1.8 Å or better and with less than 200 residues (the average number of residues is 120.4) from PISCES [61]. Their PDB codes are 2LIS, 1EP0, 1TIF, 1EB6, 1C1L, 1CCW, 2PTH, 1I6W, 1DBF, 1KPF, 1LRI, 1AAP, 1C75, 1CC8,

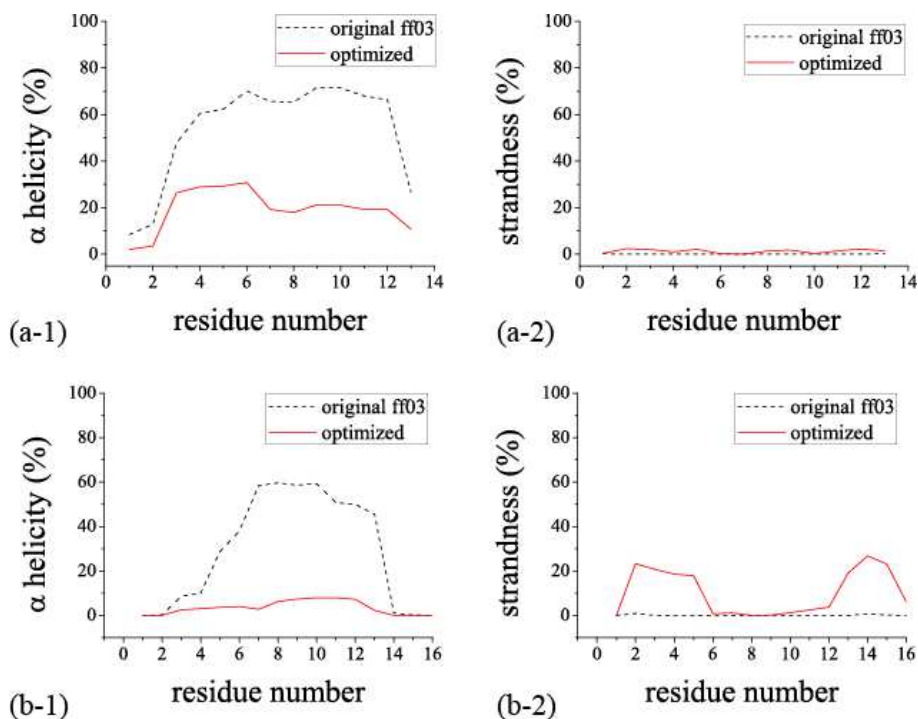


Fig. 8 α -helicity (a-1) and β -strandness (a-2) of C-peptide and α -helicity (b-1) and β -strandness (b-2) of G-peptide as functions of the residue number at 300 K. These values were obtained from the REMD simulations. Normal and dotted curves stand for the optimized and the original AMBER ff03 force fields, respectively.

1FK5, 1KQR, 1K1E, 1CZP, 1GP0, 1KOI, 1IQZ, 3EBX, 1I40, 1EJG, 1AMM, 1I07, 1GK8, 1GVP, 1M4I, 1EYV, 1E29, 1I2T, 1VCC, 1FM0, 1EXR, 1GUT, 1H4X, 1GBS, 1B0B, 1I9L, 1IFC, 1DLW, 1EAJ, 1GGZ, 1JR8, 1RB9, 1VAP, 1JZG, 1M55, 1EN2, 1C9O, 2ERL, 1EMV, 1F41, 1EW6, 2TNF, 1IFR, 1JSE, 1KAF, 1HZZ, 1HQK, 1FXL, 1BKR, 1ID0, 1LQV, 1G2R, 1KR7, 1QTN, 1D4O, 1EAZ, 2CY3, 1UGI, 1IJV, 3VUB, 1BZP, 1JYR, 1DZK, 1QFT, 1UTG, 2CPG, 1I6W, 1C7K, 1I8O, 1LO7, 1LNI, 1EQO, 1NDD, 1HD2, 3PYP, 1FD3, 1DK8, 1WHI, 1FAZ, 4FGF, 2MHR, 1JB3, 2MCM, 1IGD, 1C5E, and 1JIG.

In Step 2 of the flowchart, we used the routine in the TINKER package to add hydrogen atoms to the PDB coordinates. The force fields that we optimized are the AMBER parm94 version [7], parm96 version [8], parm99 version [9], CHARMM version 22 [12], and OPLS-AA [15]. We have optimized only two sets of parameters. The first set is the partial-charge parameters (q_i in Eqs. (5) and (27)). In order to simplify the constraint-imposing processes on the total charge, we did not optimize the charge of one of the hydrogen atoms (HN) in proline when it is located at the N-terminus. In the original X-ray data, hydrogen coordinates are missing, and in the

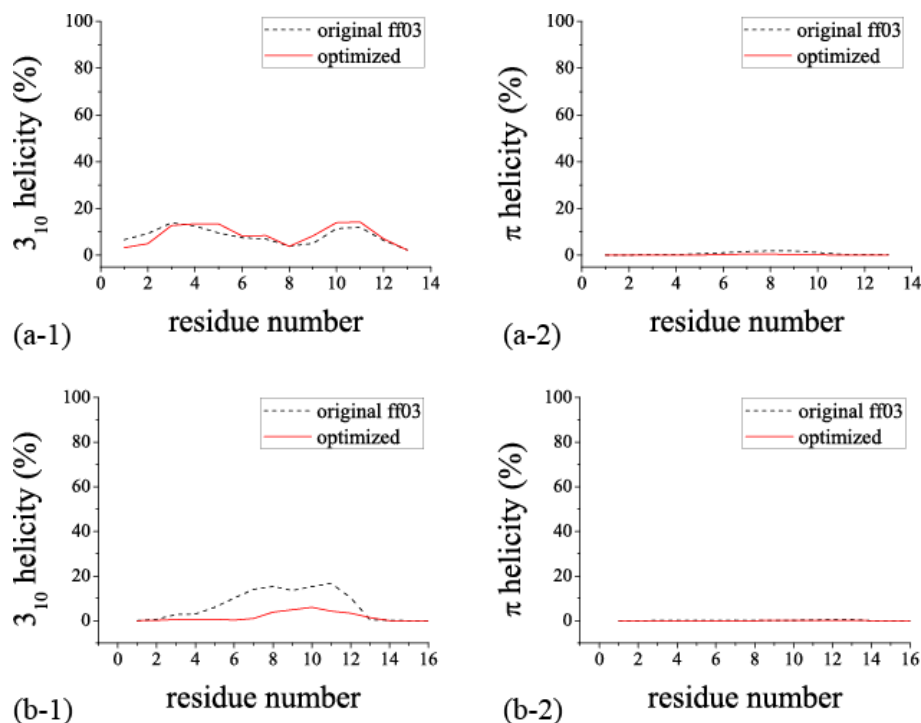


Fig. 9 3_{10} -helicity (a-1) and π -helicity (a-2) of C-peptide and 3_{10} -helicity (b-1) and π -helicity (b-2) of G-peptide as functions of the residue number at 300 K. These values were obtained from the REMD simulations. Normal and dotted curves stand for the optimized and the original AMBER ff03 force fields, respectively.

case of neutral histidine whether N_{δ} and N_{ϵ} are protonated or not is non-trivial to determine. Because we want to deal with as many as PDB data as possible, we treated all the histidine residues as positively charged histidine for simplicity. Among the five force fields, AMBER has the largest number of remaining partial-charge parameters (602). We thus optimized these 602 parameters for all the five force fields. The second set of parameters that we optimized is the backbone torsion-energy parameters (V_a , V_b , and V_c in Eq. (30)) and there are six such parameters (three each for ϕ and ψ).

As explained in detail above, the coordinates of the 100 proteins molecules have been prepared (Steps 1 and 2 of the flowchart in Fig. 1). The coordinate refinement in Step 3 of the flowchart was then carried out with the constraint in Eq. (29) on the heavy atoms. As for the force constant K_x in Eq. (29), we have some freedom for the choice of the values. Our choice is: K_x should be of the same order as K_l in the bond-stretching term in Eq. (3). The force constant K_l in AMBER varies from 166 kcal/mol/Å² to 656 kcal/mol/Å², and that in CHARMM varies

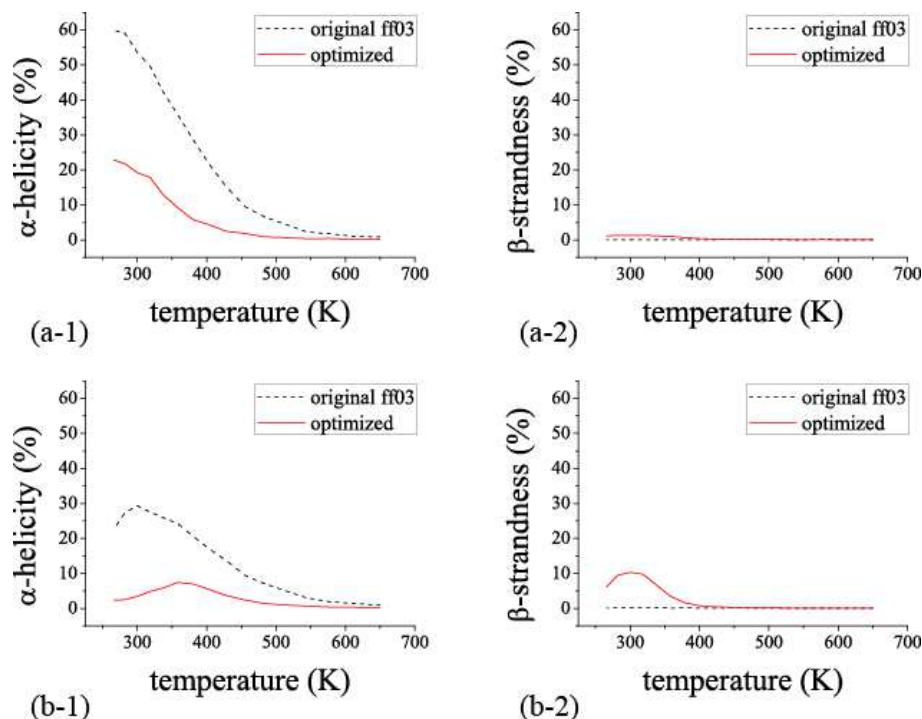


Fig. 10 α -helicity (a-1) and β -strandness (a-2) of C-peptide and α -helicity (b-1) and β -strandness (b-2) of G-peptide as functions of temperature. These values were obtained from the REMD simulations. Normal and dotted curves stand for the optimized and the original AMBER ff03 force fields, respectively.

from 173 kcal/mol/Å² to 650 kcal/mol/Å². Hence, in our first trial we set $K_x = 100$ kcal/mol/Å².

In Step 4 of the flowchart, we performed the optimization of the 602 partial-charge parameters by MC simulated annealing. Namely, we minimized F in Eq. (23) by MC simulated annealing simulations of these parameters (the parameters were updated and the updates were accepted or rejected according to the Metropolis criterion). For this we introduced an effective “temperature” for the parameter space. The simulation run consisted of 50,000 MC sweeps with the temperature decreased exponentially from 20 to 0.01. The simulation was repeated 10 times with different initial random numbers. The time series of F from these simulations are shown in Figs. 11(a)–11(e). We see that F decreases quickly in the beginning until about 5,000 MC sweeps and then it decreases very slowly for all force fields; the total number of MC sweeps (50,000) seems sufficient. The optimized partial charges are taken from those that resulted in the lowest F value.

In Tables 7–9, five examples (glycine, alanine, and glutamic acid) of the obtained partial charges together with the original force-field values are listed. We see from

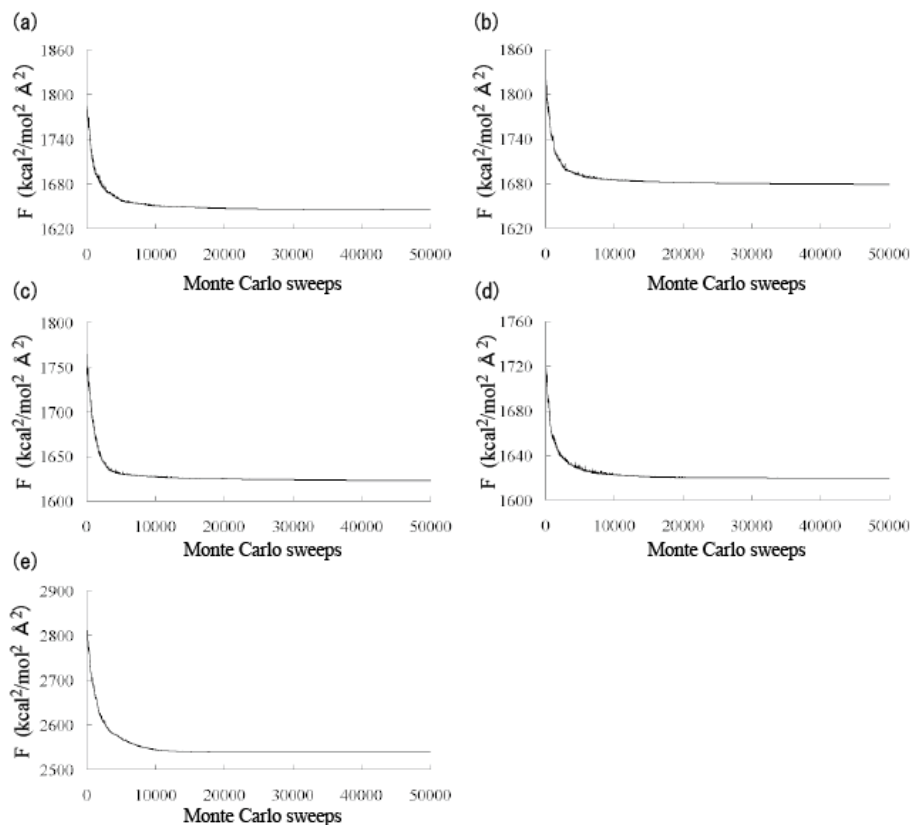


Fig. 11 Time series of MC simulated annealing simulations in force-field parameter space of partial charges for AMBER parm94 (a), AMBER parm96 (b), AMBER parm99 (c), CHARMM version 22 (d), and OPLS-AA (e). The ordinate is the value of F in Eq. (23).

these tables that the values of the partial charges have not changed a lot. Although the sign of the partial charges remains the same for those with large magnitude, charges with small magnitude sometimes change their signs (see, for example, CA of glycine and CG of glutamic acid).

In Step 5 of the flowchart, the original coordinates obtained in Step 2 were again refined with the constraints in Eq. (29), but this time the optimized parameters from Step 4 were used. This time we used the value $K_x = 500$ kcal/mol/Å². For all force fields, the average RMSD of the 100 proteins is 0.012 Å, and the coordinates of heavy atoms have little changed.

In Step 6 of the flowchart, we carried out the optimization of the six torsion-energy parameters (V_a , V_b , and V_c in Eq. (30) for both ϕ and ψ) by minimizing F in Eq. (23) with MC simulated annealing simulations in this parameter space. The simulation run consisted of 10,000 MC sweeps with the temperature decreasing

Table 7 Partial-charge parameters of glycine. AMB, CHA, and OPLS respectively stand for the original AMBER, CHARMM version 22, and OPLS-AA force fields. Opt(94), Opt(96), Opt(99), Opt(CH), and Opt(OP) are the optimized AMBER parm94, AMBER parm96, AMBER parm99, CHARMM version 22, and OPLS-AA, respectively.

[illegible]

Table 8 Partial-charge parameters of alanine. See the caption in Table 7.

[illegible]

Table 9 Partial-charge parameters of glutamic acid. See the caption in Table 7.

[illegible]

from 1,000 to 1.0. The simulation was repeated six times with different random numbers. We stopped after six trials because the convergence was very good. The optimized torsion-energy parameters are taken from those that resulted in the lowest F value. The obtained torsion-energy parameters are listed in Tables 10 and 11.

Table 10 Torsion parameters of ϕ angle. Parm94, Parm96, Parm99, CHARMM, and OPLS are AMBER parm94, AMBER parm96, AMBER parm99, CHARMM version 22, and OPLS-AA force fields, respectively. “Optimized” stands for the corresponding optimized force field.

Force field	V_a	n_a	γ_a	V_b	n_b	γ_b	V_c	n_c	γ_c
Parm94	0.200	2	180.0	—	—	—	—	—	—
Optimized	0.191	1	0.0	0.146	2	180.0	-0.223	3	0.0
Parm96	0.850	1	0.0	0.300	2	180.0	—	—	—
Optimized	1.182	1	0.0	0.359	2	180.0	-0.410	3	0.0
Parm99	0.800	1	0.0	0.850	2	180.0	—	—	—
Optimized	1.380	1	0.0	0.599	2	180.0	-0.330	3	0.0
CHARMM	0.200	1	180.0	—	—	—	—	—	—
Optimized	-0.047	1	180.0	0.240	2	180.0	-0.015	3	0.0
OPLS	-2.365	1	0.0	0.912	2	180.0	-0.850	3	0.0
Optimized	0.502	1	0.0	1.811	2	180.0	-0.567	3	0.0

Table 11 Torsion parameters of ψ angle. See the caption in Table 10.

Force field	V_a	n_a	γ_a	V_b	n_b	γ_b	V_c	n_c	γ_c
Parm94	0.750	1	180.0	1.350	2	180.0	0.400	4	180.0
Optimized	-0.368	1	180.0	1.658	2	180.0	0.265	4	180.0
Parm96	0.850	1	0.0	0.300	2	180.0	—	—	—
Optimized	0.039	1	0.0	1.011	2	180.0	0.104	3	0.0
Parm99	1.700	1	180.0	2.000	2	180.0	—	—	—
Optimized	0.228	1	180.0	1.684	2	180.0	-0.031	3	0.0
CHARMM	0.600	1	0.0	—	—	—	—	—	—
Optimized	0.321	1	0.0	0.028	2	180.0	0.251	3	0.0
OPLS	1.816	1	0.0	1.222	2	180.0	1.581	3	0.0
Optimized	0.880	1	0.0	1.479	2	180.0	0.952	3	0.0

In the present work, we stopped our process in Step 6 of the flowchart and did not iterate the optimizations.

In order to examine how much the torsion-energy terms have changed after optimizations, we depict them in Fig. 12 (we remark that the error of factor 2 in the ordinate of Fig. 5 (e1) in Ref. [26] is corrected here). Although the behaviors of the original force fields are quite different, those of the optimized force fields are rather similar. For example, the optimized torsion-energy curves for ϕ angles have two maximum peaks around $\phi \sim -60^\circ$ and $+60^\circ$ and a local minimum at $\phi = 0^\circ$, while

those for ψ angle have two peaks around $\psi \sim -100^\circ$ and $+100^\circ$ and a local minimum at $\psi = 0^\circ$ (the exceptions are those for CHARMM version 22 and OPLS-AA, which give the global maximum and a local maximum, respectively, at $\psi = 0^\circ$). These results suggest that our optimizations of the torsion-energy term yield a tendency for convergence towards a common function. Some remark is in order. The case for the optimized CHARMM is the most distinct from other optimized parameters in the sense that it gives the global maximum as $\psi = 0^\circ$ whereas that for other cases lie around $\psi \sim -100^\circ$ and $+100^\circ$.

In Fig. 13 the potential-energy surfaces of the alanine dipeptide (ACE-ALA-NME) are shown for the 10 force-field parameters: the original AMBER parm94, AMBER parm96, AMBER parm99, CHARMM version 22, OPLS-AA, and the corresponding optimized parameters. According to the *ab initio* quantum mechanical calculations, there exist three local-minimum states in the energy surface [7]. They are conformers C_{7eq} , C_5 , and C_{7ax} , which correspond to $(\phi, \psi) \sim (-80^\circ, +80^\circ)$, $(-160^\circ, +160^\circ)$, and $(+75^\circ, -60^\circ)$, respectively (C_{7eq} is the global-minimum state). We remark that these are the results of quantum chemistry calculations in vacuum, and so it is not clear how reliable the results are to represent the dipeptide in aqueous solution. The results of all five original force fields in Figs. 13(a1)–13(e1) seem to satisfy the above conditions. Namely, there are three local-minimum states at the locations of C_{7eq} , C_5 , and C_{7ax} , and the global-minimum state is C_{7eq} . As for the results of the optimized force fields in Figs. 13(a2)–13(e2), those for CHARMM version 22 and OPLS-AA also satisfy the above conditions. Those of the optimized AMBER force fields are less consistent with the quantum mechanical calculations: C_{7eq} is no longer the global-minimum state, but it is a local-minimum state. In particular, the optimized AMBER parm99 seems to be in the greatest disagreement in the sense that the C_{7eq} state is almost disappearing.

We now present another example of the refinement of our backbone torsion energy in Eq. (14). We consider the following truncated Fourier series:

$$\begin{aligned} \mathcal{E}(\phi, \psi) = & a + b_1 \cos \phi + c_1 \sin \phi + b_2 \cos 2\phi + c_2 \sin 2\phi \\ & + d_1 \cos \psi + e_1 \sin \psi + d_2 \cos 2\psi + e_2 \sin 2\psi \\ & + f_{11} \cos \phi \cos \psi + g_{11} \cos \phi \sin \psi \\ & + h_{11} \sin \phi \cos \psi + i_{11} \sin \phi \sin \psi . \end{aligned} \quad (41)$$

This function has 13 Fourier-coefficient parameters. We will see below that this number of Fourier terms is sufficient for the most of our purposes [34, 35], but that for some cases more number of Fourier terms are preferred.

We optimize the force-field parameters of this double Fourier series by using our optimization method. At first, we chose 100 PDB files with resolution 2.0 Å or better, with sequence similarity of amino acid 30.0 % or lower and with less than 200 residues (the average number of residues is 117.0) from PDB-REPRDB [55]. Generally, data from X-ray experiments do not have hydrogen atoms. Therefore, we have to add hydrogen coordinates. Many protein simulation software packages

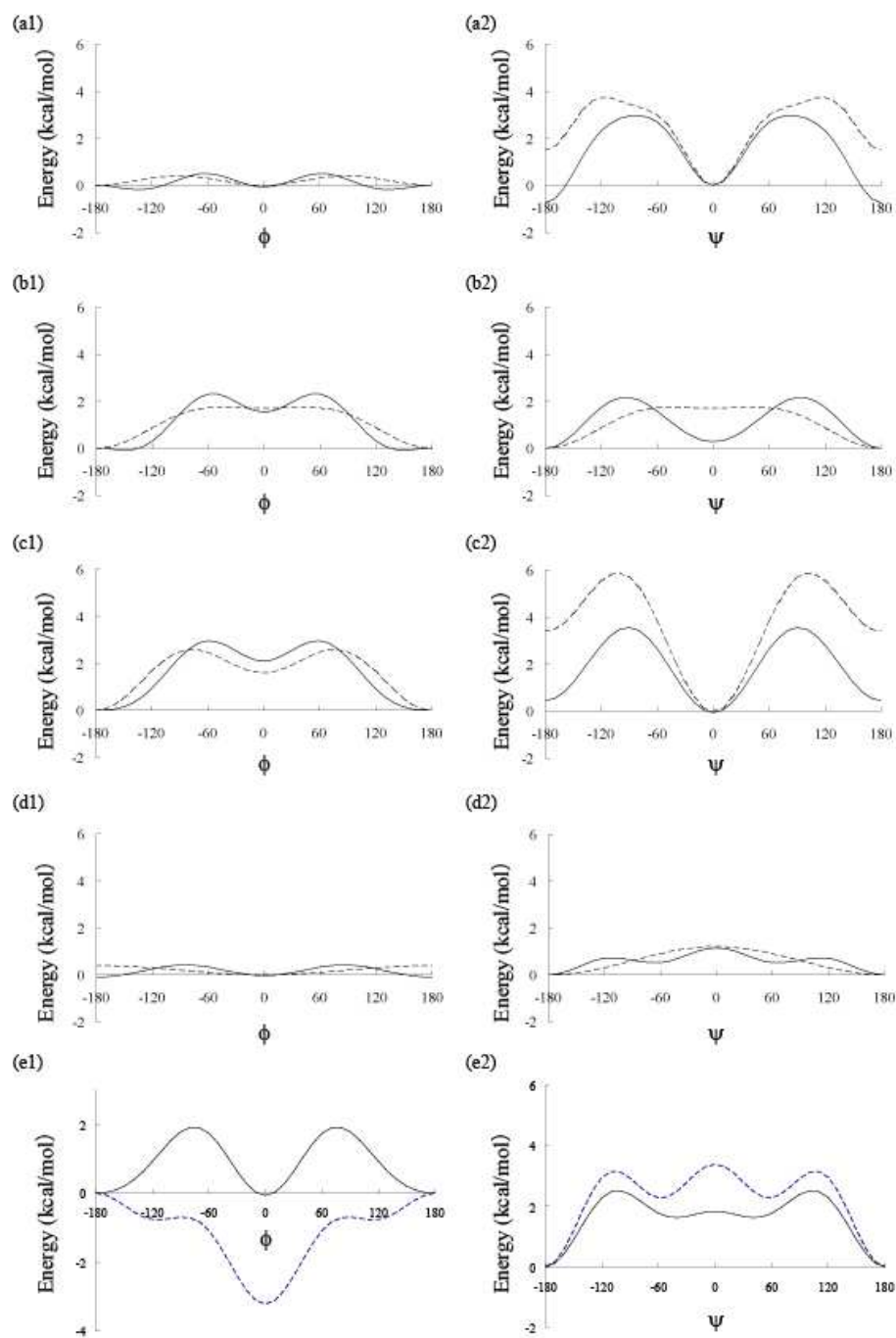


Fig. 12 Backbone torsion-energy curves as functions of ϕ (in degrees) and ψ (in degrees). The force fields are AMBER parm94 (a), AMBER parm96 (b), AMBER parm99 (c), CHARMM version 22 (d), and OPLS-AA (e). The results for the original force fields are represented by dotted curves, and those for the optimized force fields are by solid curves.

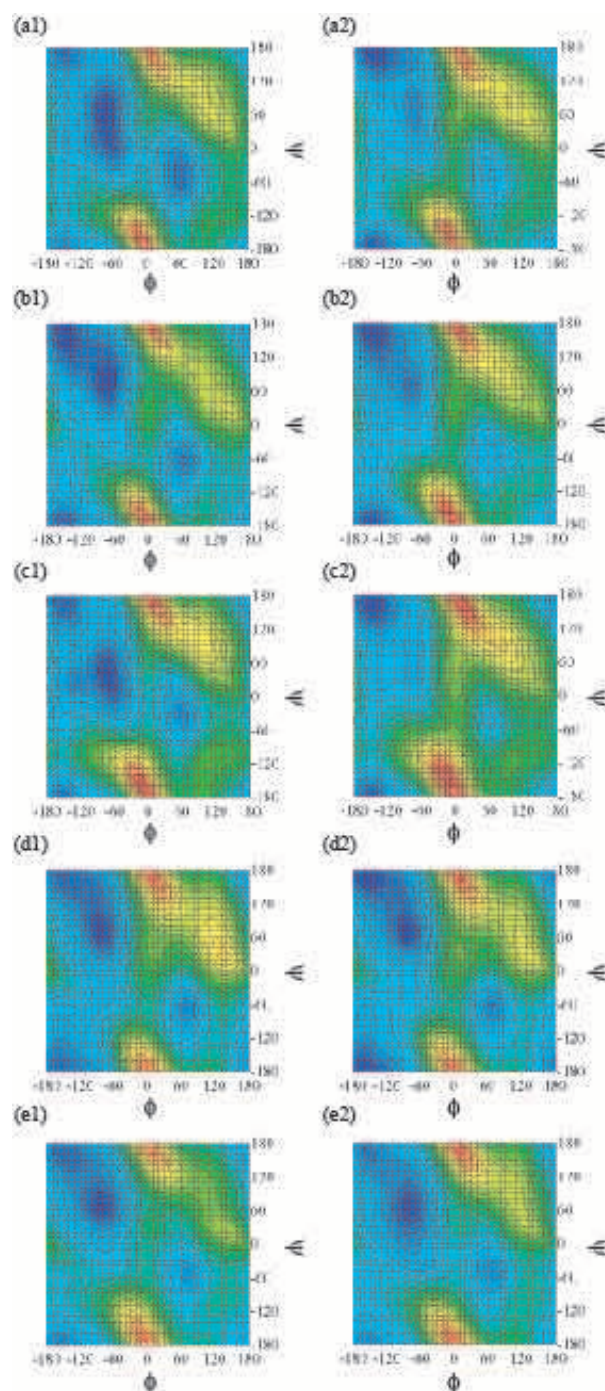


Fig. 13 Potential-energy surfaces of alanine dipeptide. The force fields are the original AMBER parm94 (a1), AMBER parm96 (b1), AMBER parm99 (c1), CHARMM version 22 (d1), and OPLS-AA (e1), and the corresponding optimized parameters (a2)-(e2). The contour maps were evaluated every 10° of ϕ and ψ angles and plotted every 1 kcal/mol, after minimizing the total potential energy in vacuum with the backbone structures fixed. The bluer the color is, the lower the potential energy surface is. As the potential-energy value increases, the color changes from blue to green, to yellow, and to red.

provide with routines that add hydrogen atoms to the PDB coordinates. We used the TINKER program package [53].

In our optimization method, the minimizations of F in Eq. (23) by the Monte Carlo (MC) simulations of the 13 backbone-torsion-energy parameters with 3000 MC steps were performed. The initial values of 13 parameters were all set to be zero. We performed MC simulations of the optimization for each f_{cut} value 10 times with different seeds for the random numbers. After that, the minimum F value was selected from the results of the obtained 10 parameter sets for each case of the f_{cut} value. The overall parameter distributions were essentially the same for the 10 runs. The maximum f_{cut} value was taken to be $f_{\text{cut}}^{\text{max}} \simeq 9.0$, which was selected from the peak point in the distribution of the forces acting on each atom in the 100 protein structures in Fig. 14. For the obtained several parameters, several ΦRMSD were calculated by using Eq. (32). Here, if a difference between Φ_i^{native} and Φ_i^{min} of a backbone dihedral angle in a protein was more than 20 degrees, the value was ignored. Because there are about 90% of differences between Φ_i^{native} and Φ_i^{min} including less than 20 degrees. In Fig. 15, the distribution of the backbone dihedral angles in the 100 protein structures is shown. Namely, we wanted to consider the majority of the differences of backbone dihedral angles. After the calculations of several ΦRMSD , we select $f_{\text{cut}} = 8.5$ at the minimum value of ΦRMSD from the several those.

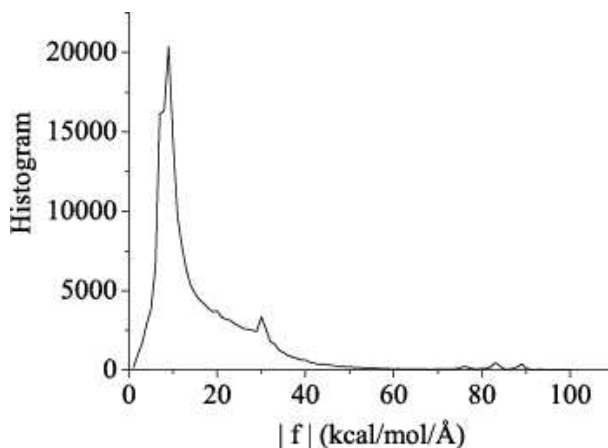


Fig. 14 The distribution of the absolute value of the forces acting on each atom in the 100 protein structures, which were obtained from PDB.

In Table. 12, optimized double Fourier-coefficient parameters and the corresponding original AMBER ff94 and ff96 force-field parameters are listed. Here, the original AMBER ff94 has a Fourier coefficient that the number of waves is four. Therefore, this coefficient set of the original AMBER ff94 is not complete. Addi-

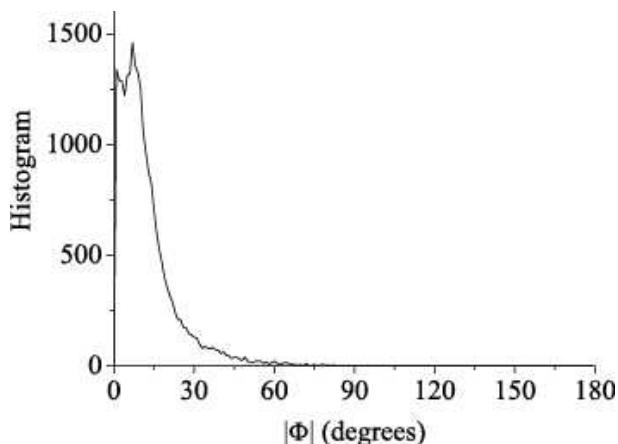


Fig. 15 The distribution of the absolute value of the backbone dihedral angles Φ (ϕ and ψ) in the 100 protein structures, which were obtained from PDB.

tionally, in Fig. 16, these backbone-torsion-energy surfaces on the Ramachandran space are illustrated.

Table 12 Fourier coefficients in Eq. (39) obtained from the numerical evaluations of the integrals in Eq. (15). “org94” and “org96” stand for the original AMBER ff94 and the original AMBER ff96, respectively, “optimized” stands for the optimized force field obtained by our optimization method. Here, the original AMBER ff94 has the Fourier coefficient that the number of waves is four. Therefore, this coefficient set of the original AMBER ff94 is not complete.

coefficient	org94	org96	optimized
a	2.700	2.300	0.000
b_1	0.000	0.850	0.835
b_2	-0.200	-0.300	-0.088
c_1	0.000	0.000	-0.327
c_2	0.000	0.000	0.100
d_1	-0.750	0.850	0.287
d_2	-1.350	-0.300	0.019
e_1	0.000	0.000	-0.160
e_2	0.000	0.000	-0.054
f_{11}	0.000	0.000	-0.427
g_{11}	0.000	0.000	0.247
h_{11}	0.000	0.000	0.114
i_{11}	0.000	0.000	0.603

In order to test the validity of the force-field parameters obtained by our optimization methods, we performed folding simulations using two peptides, namely, C-peptide and G-peptide.

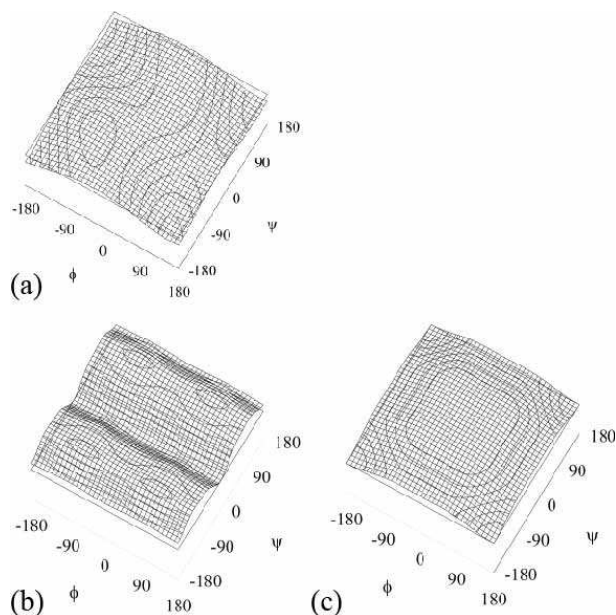


Fig. 16 The backbone-torsion-energy surfaces of the optimized force field (a), the original AMBER ff94 (b), and the original AMBER ff96 are shown.

For the folding simulations, we used the replica-exchange molecular dynamics (REMD) method [59]. We used the TINKER program package [53] modified by us for the folding simulations. The unit time step was set to 1.0 fs. Each simulation was carried out for 5.0 ns (hence, it consisted of 5,000,000 MD steps) with 32 replicas. The temperature during MD simulations was controlled by Nosé-Hoover method [62]. For each replica the temperature was distributed exponentially from 700 K to 250 K. As for solvent effects, we used the GB/SA model [41, 42] included in the TINKER program package [53].

We checked the secondary-structure formations, such as the helicity and the strandness, by using the DSSP program [44], which is based on the formations of the intra-backbone hydrogen bonds. Strandness means that there are β -bridge or extended strand in the corresponding amino acid. In Fig. 17, the helicity and strandness of C-peptide which were obtained with the optimized force field, the original AMBER ff94 and ff96 are shown. In comparison with the helicity of the original AMBER ff94, the helicity of the optimized force field decreases and in comparison with that of the original AMBER ff96, that of the optimized force field increases. For the strandness, the original AMBER ff94 is almost zero, and both the optimized force field and the original AMBER ff96 have the low strandness.

In Fig. 18, the helicity and strandness of G-peptide which were obtained with the optimized force field, the original AMBER ff94 and ff96 are shown. The helicity of the original AMBER ff94 obviously has high value the same as the case of C-

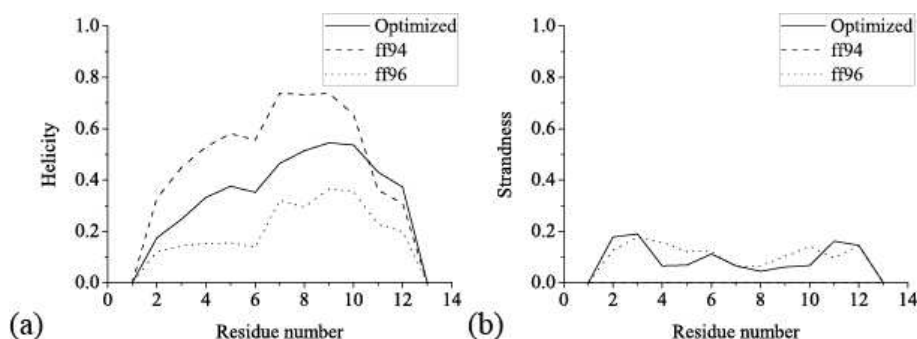


Fig. 17 Helicity (a) and strandness (b) of C-peptide as functions of the residue number. These values are obtained from the REMD [59] simulations at 300K. Normal, dashed, and dotted lines stand for the optimized force field, the original AMBER ff94, and the original AMBER ff96, respectively. There is only one secondary structural element (an α -helix in residues 4 to 12) in the native structure (PDB ID: 1A5P). See Fig. 5(a).

peptide. On the other hand, the helicity of both the optimized force field and the original AMBER ff96 decrease in comparison with the case of the original AMBER ff94. However, in comparison with the original AMBER ff96, the optimized force field slightly favors the helix structure in the region around amino-acid residues 6–8. In the experimental results, there is a turn region around residues 7–10 in G-peptide, and the backbone-torsion angles of the turn conformation are similar to that of the helix structure. Therefore, we consider that this tendency is not disagreement with the experimental results. For the strandness, the original AMBER ff94 is also almost zero the same as the case of C-peptide, and both the optimized force field and the original AMBER ff96 have higher values of the strandness than those of the helicity. In Fig. 18(b), the strandness decreases in the region around 7–8 residues in agreement with the experiments.

These secondary-structure-forming tendencies of the optimized force field for two peptides agree with experimental implications in comparison with those of the original AMBER ff94 and ff96 force field. Therefore, our improvement methods succeeded in enhancing the accuracy of the AMBER force field.

3.2.2 Use of RMSD I [38]

We now present the results of the applications of our optimization method in Subsection 2.3.2, which we refer to as Method 2, as well as that in Subsection 2.3.1, which we refer to as Method 1.

At first, we chose 100 PDB files with resolution 2.0 Å or better, with sequence similarity of amino acid 30.0 % or lower and with less than 200 residues (the average number of residues is 117.0) from PDB-REPRDB [55]. Next, we refine these selected 100 structures. Generally, data from X-ray experiments do not have hydro-

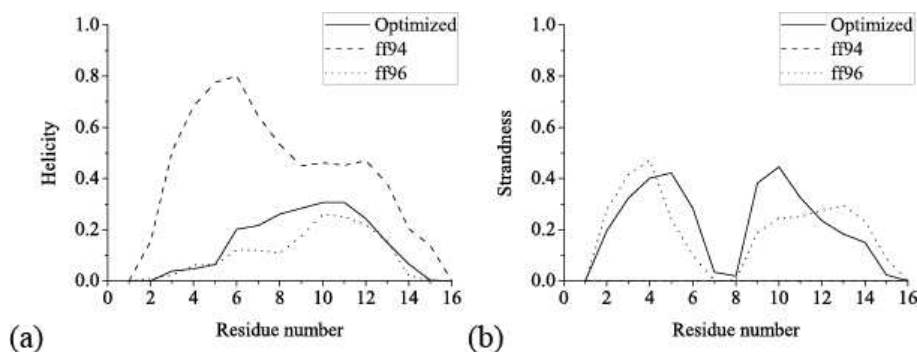


Fig. 18 Helicity (a) and strandness (b) of G-peptide as functions of the residue number. These values are obtained from the REMD [59] simulations at 300K. Normal, dashed, and dotted lines stand for the optimized force field, the original AMBER ff94, and the original AMBER ff96, respectively. There is only one secondary structural element (a β -hairpin; β -strands are in residues 2 to 6 and residues 11 to 15) in the native structure (PDB ID: 1PGA). See Fig. 5(b).

gen atoms. Therefore, we have to add hydrogen coordinates. Many protein simulation software packages provide with routines that add hydrogen atoms to the PDB coordinates. We used the TINKER program package [53]. We thus minimize the total potential energy $E_{\text{total}} = E_{\text{conf}} + E_{\text{solv}} + E_{\text{constr}}$ with respect to the coordinates for each protein conformation, where E_{constr} is the constraint energy term in Eq. (29). Here, K_x is the force constant of the restriction and \mathbf{x}_0 are the original coordinate vectors of heavy atoms in PDB. As one can see from Eq. (29), the coordinates of hydrogen atoms will be mainly adjusted, but unnatural heavy-atom coordinates will also be modified. We performed this minimization for all the 100 protein structures separately and obtained 100 refined structures.

We focused on the parameters of torsion-energy term, which we believe to be an important force-field term that influences the backbone conformational preferences such as α -helix structure and β -sheet structure. For example, AMBER parm94 [7] and AMBER parm96 [8] have very different behaviors about the secondary-structure-forming tendencies, although these force fields differ only in the backbone torsion-energy terms for rotations of the backbone ϕ and ψ angles. Recently, new force-field parameters of the backbone torsion-energy term about ϕ and ψ angles have been developed, which are, e.g., AMBER ff99SB [10], AMBER ff03 [11], and CHARMM 22/CMAP [13].

The force field that we optimized is the OPLS-UA [63]. The torsion-energy term $E_{\text{torsion}}(\Phi)$ for this force field is given by Eq. (5). We performed the force-field parameter optimizations that correspond to the following torsion angles by Methods 1 and/or 2.

1. $\text{N}-\text{C}_\alpha-\text{C}_\beta-\text{C}_\gamma$ and $\text{C}-\text{C}_\alpha-\text{C}_\beta-\text{C}_\gamma$ (χ_1) by Method 2

2. C-N-C $_{\alpha}$ -C (ϕ), N-C $_{\alpha}$ -C-N (ψ), C-N-C $_{\alpha}$ -C $_{\beta}$ and N-C-C $_{\alpha}$ -C $_{\beta}$ by Methods 1 and 2
3. C-N-C $_{\alpha}$ -C $_{\beta}$ by Method 2
4. N-C $_{\alpha}$ -C-N by Method 2
5. C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$ -C $_{\delta}$ (χ_2 of Glu) by Methods 1 and 2

Here, we also optimized the force-field parameters of χ_2 of Glu. The reason is given below.

In Method 1, the minimizations of F in Eq. (23) by the Monte Carlo (MC) simulated annealing simulations of the torsion-energy parameters with 10000 MC steps were performed 10 times. Here, we neglected the improper-torsion-energy contributions to E_{conf} in Eq. (25). In order to make a better force field, we have to optimize many force-field parameters. However, we ignored the uncertainty of improper-torsion-energy parameters with this optimization, because we wanted to focus on the torsion-energy parameters and Method 1 is very sensitive for the energy of dihedral angles. For example, one of the results of the simulations of Method 1 above is shown in Fig. 19.

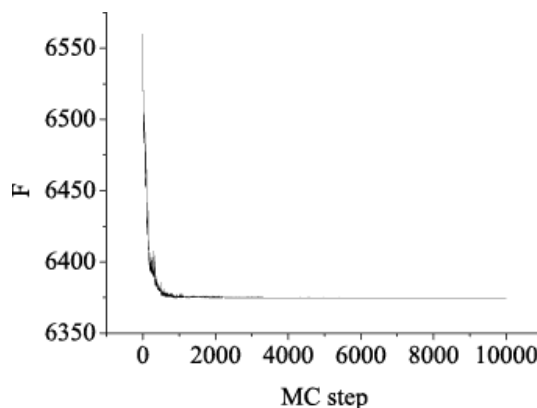


Fig. 19 Time series of Monte Carlo simulated annealing simulations in force-field parameter space of torsion-energy for OPLS-UA. The ordinate is the value of F in Eq. (23).

In Method 2, the lowest R value was selected from about 10–30 optimization runs with different initial conditions. In order to calculate R , the minimizations of 100 proteins were performed using these new parameter sets. In Table 13, all the optimized torsion-energy parameters are listed. As one can see in Table 13, the original parameters of OPLS-UA force field for the optimization are almost zero.

Table 13 Original and optimized torsion-energy parameters of OPLS-UA.

	$V_1/2$		γ_1	$V_2/2$		γ_2	$V_3/2$		γ_3
	org	opt		org	opt		org	opt	
N-C α -C β -C γ (χ_1)							0.5 or 1.0	1.950	0.0
C-C α -C β -C γ (χ_1)							0.5 or 1.0	1.950	0.0
C-N-C α -C (ϕ)	0.0	-0.662	0.0	0.0	0.277	π	0.0	-0.050	0.0
N-C α -C-N (ψ)	0.0	0.974	0.0	0.0	0.576	π	0.0	-0.083	0.0
C-N-C α -C β	0.0	0.811	0.0	0.0	0.328	π	0.0	0.155	0.0
N-C-C α -C β	0.0	0.215	0.0	0.0	0.036	π	0.0	0.015	0.0
C α -C β -C γ -C δ (χ_2 of Glu)	0.0	0.565	0.0	0.0	0.177	π	2.0	-0.025	0.0

In comparison with Method 1, Method 2 can optimize force-field parameters appropriately even if there are some errors in PDB structures. However, the computational cost of Method 2 is much larger than that of Method 1. Therefore, we could not apply Method 2 to the global optimization in the force-field-parameter space. The force-field parameters of the backbone-torsion angles need the global optimization, because we consider that these parameters are the most problematic. Thus, at first, we performed the global optimization of the backbone-torsion parameters by using Method 1. After that, Method 2 was applied only on the local region of the parameter space, which was identified as relevant by Method 1.

In order to test the validity of the force-field parameters obtained by our optimization methods, we performed folding simulations using two peptides, namely, C-peptide and G-peptide.

Only Glu amino acid appears twice in each of the two peptides. Therefore, we consider that Glu amino acid is the most important, and the χ_2 parameters were optimized for this amino acid. (Of course, we expect that it becomes a better force field if the remaining force-field parameters of other amino acids are also optimized.)

For the folding simulations, we used the replica-exchange molecular dynamics (REMD) method [59]. REMD is one of the generalized-ensemble simulation algorithms and has high conformational sampling efficiency by allowing configurations to heat up and cool down while maintaining proper Boltzmann distributions. We used the TINKER program package [53] modified by us for the folding simulations. The unit time step was set to 1.0 fs. Each simulation was carried out for 10 ns (hence, it consisted of 10,000,000 MD steps) with 16 replicas. The temperature during MD simulations was controlled by Nosé-Hoover method [62]. For each replica the temperature was distributed exponentially: 700, 662, 625, 591, 558, 528, 499, 471, 446, 421, 398, 376, 355, 336, 317, and 300 K. As for solvent effects, we used the GB/SA model [41, 42] included in the TINKER program package [53]. These folding simulations were repeated 10 times with different sets of randomly generated initial velocities.

In Fig. 20, the helicity and strandness of C-peptide which were obtained with the original OPLS-UA and its optimized force field are shown. These values are the averages of the 10 REMD simulations at 300 K. In comparison with the helicity of the original OPLS-UA, the helicity of the optimized force field increases at the

amino-acid sequence between 6 and 12. For the strandness, both the original and optimized OPLS-UA force fields are almost zero.

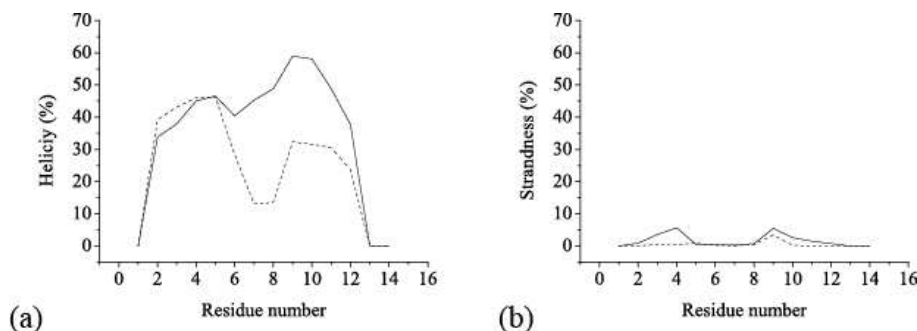


Fig. 20 Helicity (a) and strandness (b) of C-peptide as functions of the residue number. These values are the average of the 10 independent REMD [59] simulations at 300 K. Normal and dotted lines stand for the optimized and original OPLS-UA force fields, respectively.

In Fig. 21, the helicity and strandness of G-peptide at the original OPLS-UA and its optimized force field are shown. In comparison with the helicity of the original OPLS-UA, the helicity of the optimized force field decreases at the area of amino-acid sequence between 8 and 15, and in comparison with the strandness of the original OPLS-UA, the strandness of the optimized force field clearly increases at the two areas of amino-acid sequences 2–6 and 9–15. We checked the secondary-structure formations by using the DSSP program [44], which is based on the formations of the intra-backbone hydrogen bonds. Strandness means that there are β -bridge or extended strand in the corresponding amino acid. In the experimental results, there is a turn region around residues 7–10 and there are five intra-backbone hydrogen bond pairs, namely, between residue pairs 2–15, 3–14, 4–13, 5–12, and 6–11 in G-peptide. In Fig. 21(b), the strandness decreases in the region around 7–8 residues in agreement with the experiments.

These results show that the optimized force field favors helix structures more than the original OPLS-UA in the case of C-peptide and favors β structures more than the original OPLS-UA in the case of G-peptide. We see that these secondary-structure-forming-tendencies of the optimized force field are better than those of the original OPLS-UA, because these results are consistent with the native structures of the two peptides.

In Figs. 22 and 23, we show the 20 lowest-energy conformations of C-peptide and G-peptide obtained by the REMD simulations in the case of the original and optimized OPLS-UA force fields, respectively. In Fig. 22(a), five conformations (Nos. 11, 13, 16, 18, and 19) have α -helix structures for the original OPLS-UA in the case of C-peptide. In Fig. 22(b), 18 conformations (all conformations except for Nos. 2 and 12) have α -helix structures for the optimized OPLS-UA in the case of C-peptide. From these results, we can see that the optimized OPLS-UA force

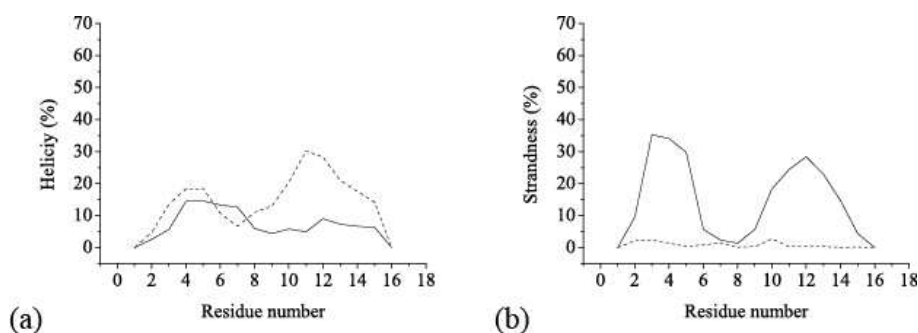


Fig. 21 Helicity (a) and strandness (b) of G-peptide as functions of the residue number. These values are the average of the 10 independent REMD [59] simulations at 300 K. Normal and dotted lines stand for the optimized and original OPLS-UA force fields, respectively.

field favor α -helix structure more than the original OPLS-UA force field in the case of C-peptide. In Fig. 23(a), 11 conformations have α -helix structures for the original OPLS-UA in the case of G-peptide. In Fig. 23(b), seven conformations have α -helix structures, and eight conformations have β -hairpin structures for the optimized OPLS-UA in the case of G-peptide. In Fig. 23(b), two conformations (Nos. 3 and 16) out of the eight β -hairpin conformations have the right hydrogen bond formations that are inferred by the experiments. Namely, conformation No.3 has three native-like hydrogen bonds between residue pairs 3–14, 4–13, and 5–12, and conformation No.16 has two native-like hydrogen bonds between residue pairs 3–14 and 4–13. These results for G-peptide show that the optimized OPLS-UA force field does not favor α -helix structure and clearly favors β -hairpin structure more than the original OPLS-UA force field.

These secondary-structure-forming tendencies of the optimized OPLS-UA force field for two peptides agree with experimental implications in comparison with those of the original OPLS-UA force field. Therefore, our optimization methods succeeded in enhancing the accuracy of the OPLS-UA force field.

3.2.3 Use of RMSD II [37]

We now present the results of the applications of our new optimization method of force-field parameters.

At first, we chose 100 PDB files with resolution 2.0 Å or better, with sequence similarity of amino acid 30.0 % or lower, and with less than 200 residues (the average number of residues is 122.2) from PDB-REPRDB [55]. We selected the number of each fold (all α , all β , α/β , and $\alpha + \beta$) in 100 proteins based on the number of folds given by SCOP (version 1.73 at November 2007) [64]. Namely, we used 29 all α , 18 all β , 16 α/β , and 37 ($\alpha + \beta$) proteins (the list is slightly different from that in Table 5).

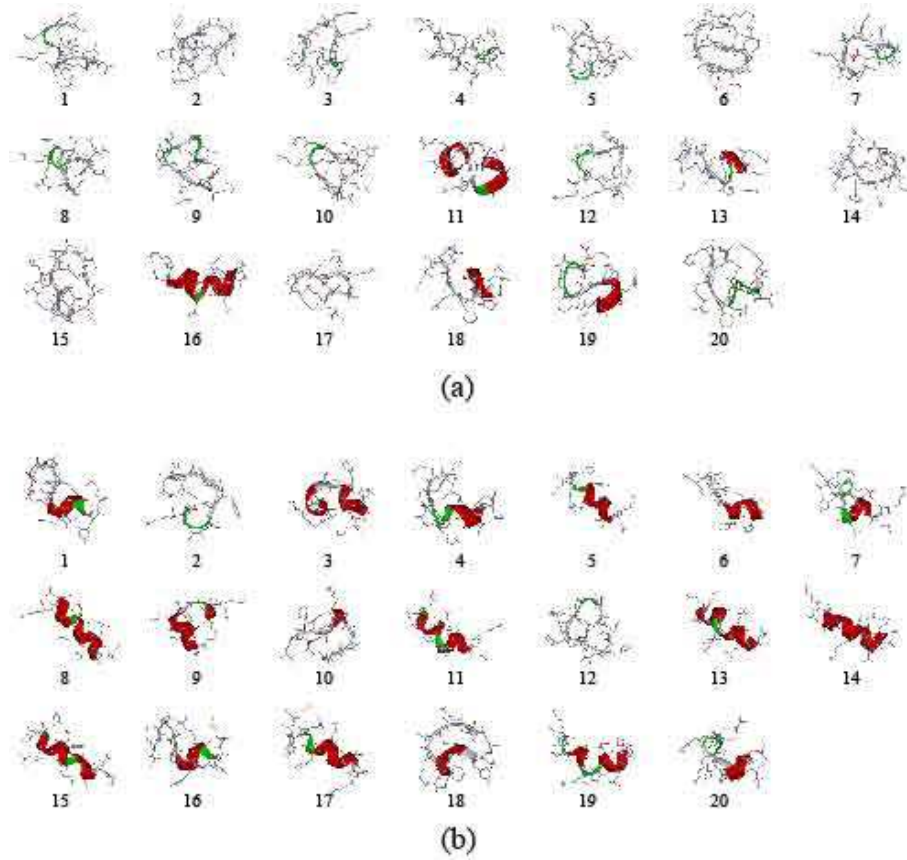


Fig. 22 Twenty lowest-energy conformations of C-peptide obtained from 10 sets of REMD [59] simulation runs. (a) and (b) are the results of the original and optimized OPLS-UA force field, respectively. The conformations are ordered in the increasing order of energy for each case. The figures were created with DS Visualizer v1.5[51].

The force field that we optimized is the AMBER parm96 version [8]. The backbone-torsion-energy term $E_{\text{torsion}}(\Phi, \Psi)$ for this force field is given by

$$E_{\text{torsion}}(\Phi, \Psi) = \frac{V_1^\phi}{2}[1 + \cos \phi] + \frac{V_2^\phi}{2}[1 - \cos 2\phi] + \frac{V_1^\psi}{2}[1 + \cos \psi] + \frac{V_2^\psi}{2}[1 - \cos 2\psi], \quad (42)$$

where we have $V_1^\phi = 1.7$, $V_2^\phi = 0.6$, $V_1^\psi = 1.7$, and $V_2^\psi = 0.6$. Here, we have optimized only two parameters in the backbone-torsion-energy term, namely, V_1^ψ and V_2^ψ for ψ angle. As described above, AMBER parm94 and AMBER parm96 have quite different secondary-structure-forming-tendencies, although these force fields differ only in the backbone torsion-energy terms for rotations of the ϕ and ψ angles. Moreover, we can easily imagine that force-field parameters V_1^ψ and V_2^ψ for ψ angle

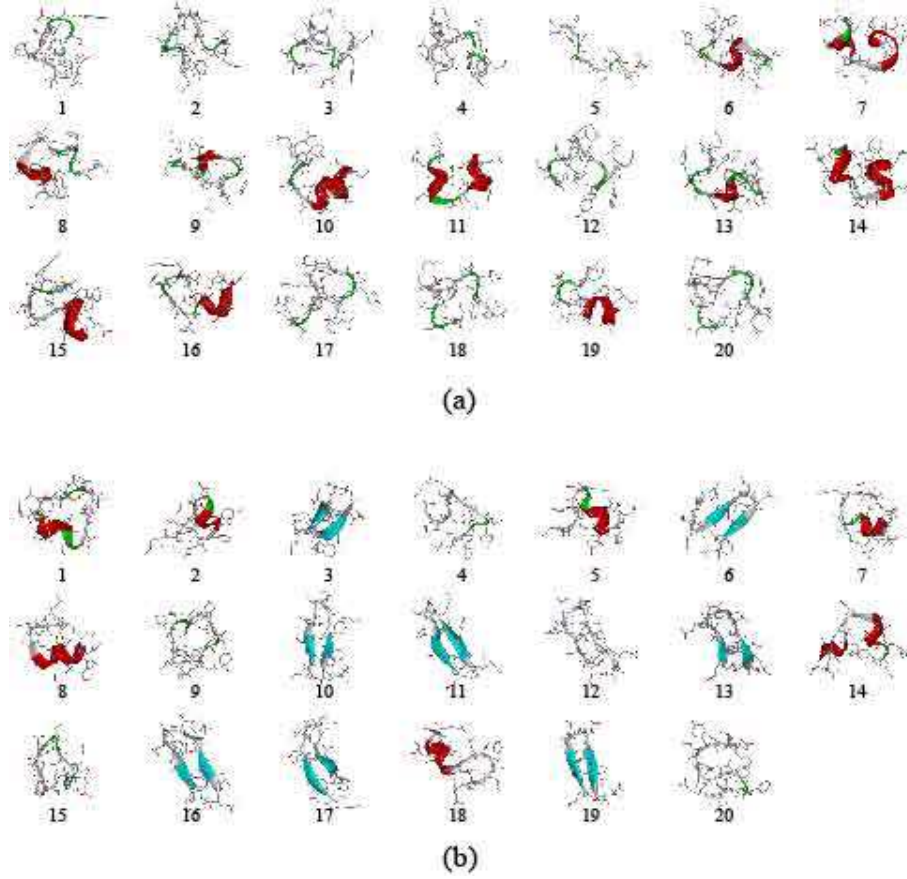


Fig. 23 Twenty lowest-energy conformations of G-peptide obtained from 10 sets of REMD [59] simulation runs. (a) and (b) are the results of the original and optimized OPLS-UA force field, respectively. The conformations are ordered in the increasing order of energy for each case. The figures were created with DS Visualizer v1.5[51].

are important for the secondary-structure-forming-tendencies, because the energy surface in the Ramachandran space is quite sensitive to this energy term in the helix and β -sheet regions. Namely, if the torsion-energy term for the ψ angle changes, the stabilities of helix structure region and β -sheet region on the Ramachandran space change. Therefore, we considered some trial force-field parameters for V_1^ψ and V_2^ψ , which are given by the following equations:

$$V_1^{\text{trial}} = 1.7 \cdot 0.2i = 0.34i, \quad (43)$$

$$V_2^{\text{trial}} = 0.6 \cdot 0.2i = 0.12i. \quad (44)$$

Here, i is any real number. When i is 5, the force-field parameters V_1^{trial} and V_2^{trial} of ψ angle are equal to those of the original AMBER parm96. From our experience, if i has a small number ($i < 5$), the force field favors helix structure, and if i has a large number ($i > 5$), the force field favors β -sheet structure (see also Figs. 24 and 25 below). We calculated $\Phi\text{RMSD}_{2\text{ndly}}$ values in Eq. (37) about some trial force-field parameters obtained by changing i in Eqs. (43) and (44).

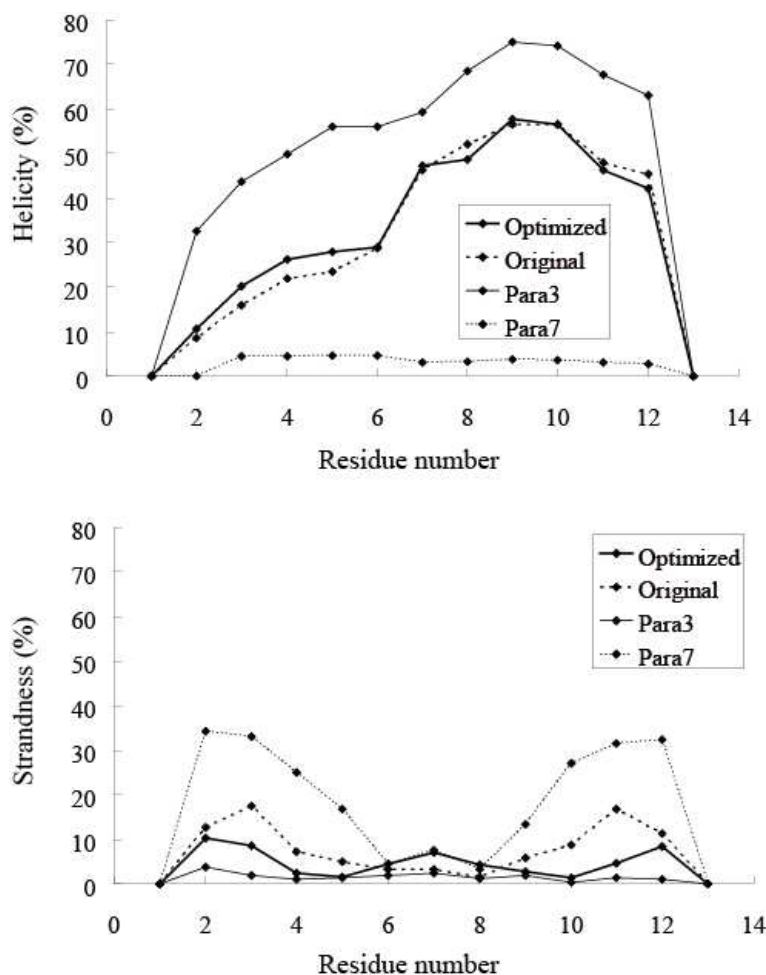


Fig. 24 Helicity (a) and strandness (b) of C-peptide as functions of the residue number. These values are the averages of the 10 independent REMD [59] simulations at 300 K. Optimized, original, para3, and para7 stand for the optimized AMBER parm96 ($i = 4.7$), original AMBER parm96 ($i = 5.0$), trial force field para3 ($i = 3.0$), and trial force field para7 ($i = 7.0$), respectively.

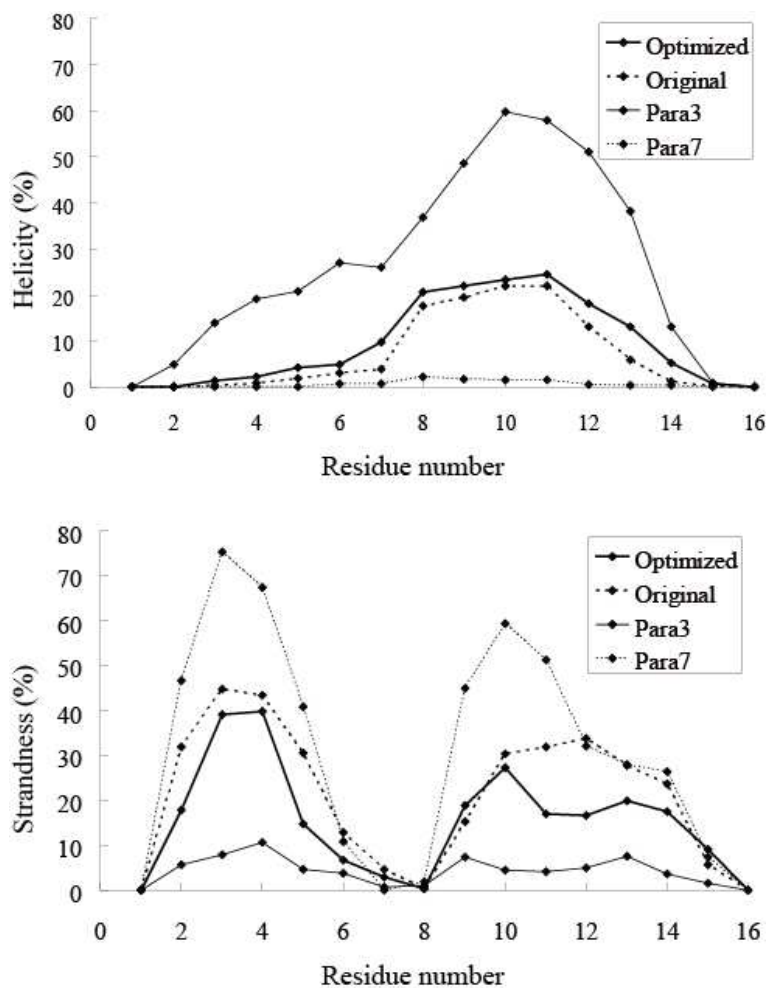


Fig. 25 Helicity (a) and strandness (b) of G-peptide as functions of the residue number. These values are the averages of the 10 REMD [59] simulations at 300 K. Optimized, original, para3, and para7 stand for the optimized AMBER parm96 ($i = 4.7$), original AMBER parm96 ($i = 5.0$), trial force field para3 ($i = 3.0$), and trial force field para7 ($i = 7.0$), respectively.

We performed the minimization, which was terminated when the root-mean-square (RMS) potential energy gradients were less than 0.1 (kcal/mol/Å) by using TINKER program package [53]. For solvent effects, we used GB/SA solvent model in TINKER.

The results of $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} are shown in Fig. 26(a) and Fig. 26(b), respectively. In these calculations, if the differences of the backbone-dihedral angles between Φ_i^{native} and Φ_i^{min} in Eq. (36) are more than 30 degrees, they were

ignored, assuming that the uncertainties in those angles are too large. We see that $\Phi\text{RMSD}_{\text{helix}}$ decreases gradually with a decrease in i . If i decreases, the torsion energy of the helix structure region in the Ramachandran space also decreases. On the other hand, ΦRMSD_{β} decreases gradually with an increase in i . If i increases, the torsion energy of the β structure region in the Ramachandran space decreases. Hence, this result is reasonable. However, ΦRMSD_{β} reaches the global minimum, when i is 6.5. If i is larger than 6.5, ΦRMSD_{β} increases gradually. This result implies that the ΦRMSD_{β} does not correspond to the parameters V_1^{trial} and V_2^{trial} completely.

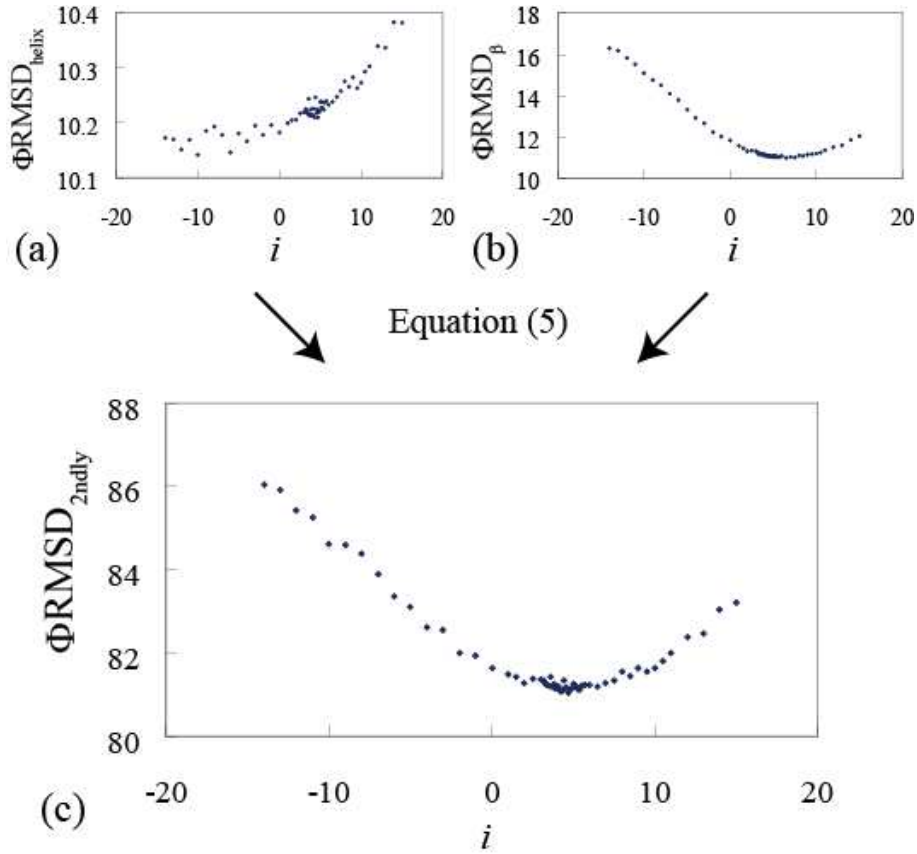


Fig. 26 Distributions of $\Phi\text{RMSD}_{\text{helix}}$ (a), ΦRMSD_{β} (b), and $\Phi\text{RMSD}_{2\text{ndly}}$ (c) obtained from the minimization of 100 proteins using the trial force-field parameters V_1^{trial} and V_2^{trial} depending on the number i .

For $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} in Fig. 26 (a) and (b), we can see the difference clearly. The noteworthy point obtained from these results is that ΦRMSD can distinguish between helix structure and β structure.

We combined $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} by Eq. (37). Here, in order to have roughly equal contributions from both terms, we can set the value of the scaling factor λ to be, for example, the coefficients of variations:

$$\lambda = \frac{\frac{\sigma_{\beta}}{\mu_{\beta}}}{\frac{\sigma_{\text{helix}}}{\mu_{\text{helix}}}}. \quad (45)$$

Here, μ_{helix} and μ_{β} are the averages and σ_{helix} and σ_{β} are the corresponding standard deviations for $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} . For the calculations, we have chosen a small number of i values in a range $i_{\min} \leq i \leq i_{\max}$. For $i_{\min} = 0$ and $i_{\max} = 10$, we obtained $\lambda = 6.857$, and this fixed value was used for all the calculations in the present work.

In Fig. 26(c), the combined result is shown. The smallest $\Phi\text{RMSD}_{2\text{ndly}}$ is obtained value $i = 4.7$, namely, the obtained force-field parameters are $V_1^{\text{trial}} = 1.598$ and $V_2^{\text{trial}} = 0.564$. These values are slightly smaller than those of the original AMBER parm96, which corresponds to $i = 5$. We can easily expect the new obtained force-field parameters slightly favor helix structure more and β -sheet structure less than the original AMBER parm96.

In order to check the force-field parameters obtained by our optimization method, we performed the folding simulations using two peptides, namely, C-peptide and G-peptide.

For the folding simulations, we used replica-exchange molecular dynamics (REMD) [59]. We used the TINKER program package [53] modified by us for the folding simulations. The unit time step was set to 1.0 fs. Each simulation was carried out for 2 ns (hence, it consisted of 2,000,000 MD steps) with 16 replicas and repeated 10 times. The temperature during MD simulations was controlled by Berendsen's method [52]. For each replica the temperature was distributed exponentially: 700, 662, 625, 591, 558, 528, 499, 471, 446, 421, 398, 376, 355, 336, 317, and 300 K. As for solvent effects, we used the GB/SA model [41, 42] included in the TINKER program package [53]. These folding simulations were performed with different sets of randomly generated initial velocities.

In Fig. 24, the helicity and strandness of C-peptide which were obtained with the original AMBER parm96 and its optimized force field are shown. These values are the averages of the 10 REMD simulations at 300 K. In comparison with the helicity of the original AMBER parm96, the helicity of the optimized force field is similler. However, the helicity of Thr3, Ala4, and Ala5 of the optimized force field slightly increases. In comparison with the strandness of the original AMBER parm96, the strandness of the optimized force field decreases except for those at Ala6, Lys7, and Phe8.

In Fig. 25, the helicity and strandness of G-peptide at the original AMBER parm96 and its optimized force field are shown. In comparison with the helicity of the original AMBER parm96, the helicity of the optimized force field slightly increases, and in comparison with the strandness of the original AMBER parm96, the strandness of the optimized force field slightly decreases. For trial force fields of para3 and para7, the secondary-structure-forming-tendencies are similar to the case of C-peptide.

These results clearly show that the optimized force field favors helix structures and does not favor β structures than the original AMBER parm96. We can see that these secondary-structure-forming-tendencies of the optimized force field are better than those of the original AMBER parm96, because it is known that the AMBER parm96 slightly favors the β structure too much [23, 24, 25, 26, 27].

We also performed the folding simulations with two extreme cases of the trial force fields, namely, para3 ($i = 3.0$) and para7 ($i = 7.0$) (see Figs. 24 and 25) for comparisons. The trial force field para3 favors helix structure strongly and does not favor β structure clearly. On the other hand, the trial force field para7 has the tendency that is quite reverse to para3. According to the results of $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} in Fig. 26(a)(b), $\Phi\text{RMSD}_{\text{helix}}$ decreases gradually with a decrease in i , and ΦRMSD_{β} reaches the global minimum, when i is 6.5. Namely, we can see that the values of $\Phi\text{RMSD}_{\text{helix}}$ and ΦRMSD_{β} are related to the stabilities of helix structure and β structure well.

3.2.4 Use of short MD simulations [45]

We present the results of the applications of our optimization method in Subsection 2.3.4 to the AMBER ff99SB force field. At first, we chose 31 PDB files ($M = 31$) with resolution 2.0 Å or better, with sequence similarity of amino acid 30.0 % or lower and with from 40 to 111 residues (the average number of residues is 86.7) from PDB-REPRDB [55]. Namely, the PDB IDs of these 31 proteins are 1LDD, 1HBK, 1Y02, 1I2T, 1U84, 2ERL, 1TQG, 1O82, 1V54, 1XAK, 1GMU, 1O5U, 1NLQ, 1WHO, 1CQY, 1H75, 1GMX, 1IIB, 1VC1, 1AY7, 1KAF, 1KPF, 1BM8, 1MK0, 1EW4, 1OSD, 1VCC, 1OPD, 1CYO, 1CTF, and 1N9L. Generally, data from X-ray experiments do not have hydrogen atoms. Therefore, we have to add hydrogen coordinates. Many protein simulation software packages provide with routines that add hydrogen atoms to the PDB coordinates. After adding the hydrogen atoms, we performed the short potential energy minimizations while restraining the heavy atoms. We use the obtained conformations as the initial structures (experimental structures). We performed MD simulations for these proteins. Each simulation was carried out for 40.0 ps (hence, it consisted of 20,000 MD steps, and the unit time step was set to 2.0 fs and the bonds involving hydrogen were constrained by SHAKE algorithm [60]) by using Langevin dynamics at 300 K. The nonbonded cutoff of 20 Å were used. As for solvent effects, we used the GB/SA model [57] included in the AMBER program package ($igb = 5$). These simulations were performed with different sets of the same generated initial velocities of atoms in 31 proteins. For

all the process, we used the AMBER11 program package [56]. As trial force-field parameters, we used the parameters V_1 of ψ (N-C α -C-N) and ψ' (C β -C α -C-N) angles for torsion-energy term in Eq. (5). We performed the simulations by using 14 and 15 values of the V_1 parameters of ψ and ψ' , respectively, and these simulations with each set of parameter values were performed five times by changing the initial velocities of atoms in the 31 proteins. Namely, we calculated $n_i^{S \rightarrow U}$ and $n_i^{U \rightarrow S}$ in Eq. (38) as the average numbers of $n_i^{S \rightarrow U}$ and $n_i^{U \rightarrow S}$ of 10 trajectories from 20.0 ps to 40.0 ps of the five simulations. These results are shown in Fig. 27. We determined the optimized force-field parameters in order of ψ' and ψ , by searching the minimum value of S in Fig. 27. V_1 parameter for ψ changed from 0.45 to 0.31, and V_1 parameter for ψ' changed from 0.20 to -1.60 .

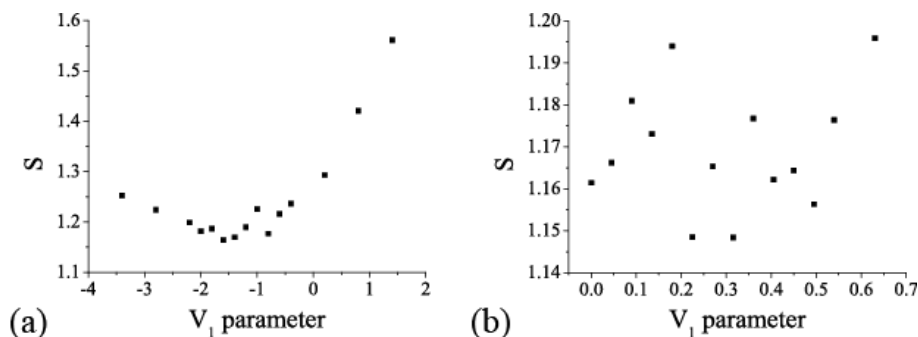


Fig. 27 S values (defined in Eq. (38)) obtained from MD simulations of 31 proteins with the force fields which have different V_1 parameter values for ψ' (C β -C α -C-N) (a) and ψ (N-C α -C-N) (b) angles.

In order to test the validity of the force-field parameters obtained by our optimization method, we performed the folding simulations using two peptides, namely, C-peptide and G-peptide.

For test simulations, we used replica-exchange molecular dynamics (REMD) [59]. We used the AMBER11 program package [56]. The unit time step was set to 2.0 fs, and the bonds involving hydrogen were constrained by SHAKE algorithm [60]. Each simulation was carried out for 30.0 ns (hence, it consisted of 15,000,000 MD steps) with 32 replicas by using Langevin dynamics. The replica exchange was tried every 3,000 steps. The temperature was distributed exponentially: 600, 585, 571, 557, 544, 530, 517, 505, 492, 480, 469, 457, 446, 435, 425, 414, 404, 394, 385, 375, 366, 357, 348, 340, 332, 324, 316, 308, 300, 293, 286, and 279 K. As for solvent effects, we used the GB/SA model [57] included in the AMBER program package ($igb = 5$). These simulations were performed with different sets of randomly generated initial velocities.

In Fig. 28, α helicity and strandness of two peptides obtained from the test simulations are shown. We checked the secondary-structure formations by using the DSSP program [44], which is based on the formations of the intra-backbone hy-

drogen bonds. For the original AMBER ff99SB force field, the α helicity is clearly larger than the strandness in not only C-peptide but also G-peptide. Namely, the original AMBER ff99SB force field clearly favors α -helix structure, and does not favor β structure. On the other hand, for the optimized force field, in the case of C-peptide, the α helicity is larger than the strandness, and in the case of G-peptide, the strandness is larger than the α helicity. We can see that these results obtained from the optimized force field are in better agreement with the experimental results in comparison with the original force field.

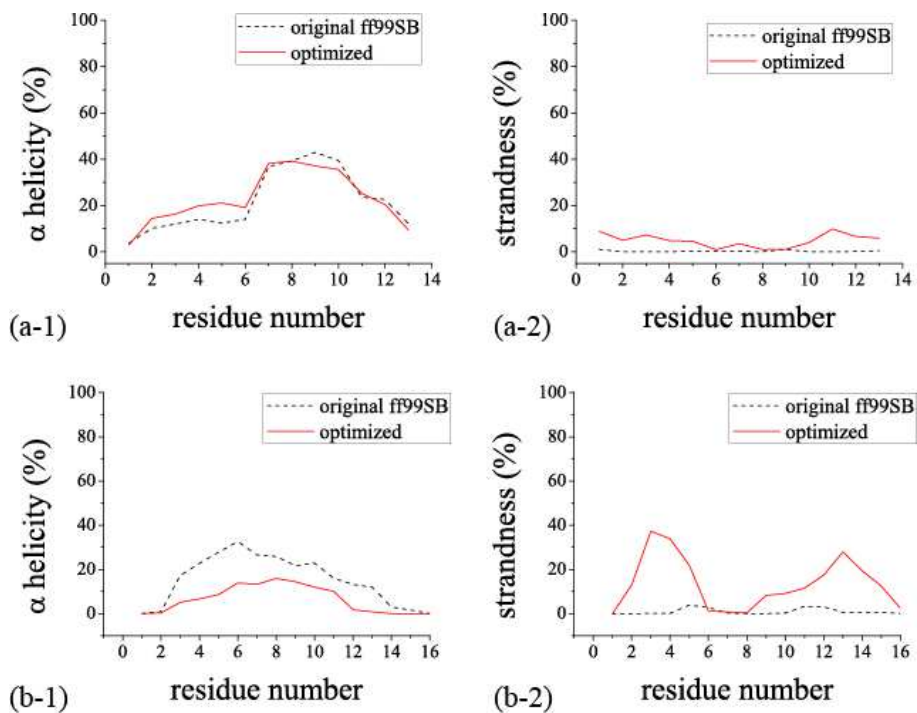


Fig. 28 α helicity (a-1) and strandness (a-2) of C-peptide and α helicity (b-1) and strandness (b-2) of G-peptide as functions of the residue number. These values are obtained from REMD [59] simulations at 300 K. Normal and dotted lines stand for the optimized and original AMBER ff99SB force field, respectively.

4 Conclusions

In this Chapter we reviewed our works on force fields for molecular simulations of protein systems. We first discussed the functional forms of the force fields and

present some extensions of the conventional ones. Because the main-chain torsion-energy terms are the most problematic among the force-field terms in the existing force fields, we mainly considered the main-chain torsion-energy terms. We have generalized them into the double Fourier series in ϕ and ψ . We have also introduced the amino-acid dependence on these terms.

Given the functional forms, we then presented various methods for force-field parameter optimizations. Some of our methods use the coordinates from PDB, which were determined by experiments. We tried to minimize the effects of systematic experimental errors by considering many protein structures. Other methods rely on short molecular dynamics simulations with the native conformations from PDB as initial ones for the simulations.

Some examples of our applications of these parameter optimization methods were given and they were compared with the results from the existing force-fields. It turned out that all the examples resulted in improvement of the existing force fields. We thus believe that we are at least on the right track.

Our optimization methods for the force-field parameters are quite general and they can be readily applied to any new energy terms whenever they are introduced in the future.

Acknowledgements The computations were performed on the computers at the Research Center for Computational Science, Institute for Molecular Science, Information Technology Center, Nagoya University, and Center for Computational Sciences, University of Tsukuba. This work was supported, in part, by the Grants-in-Aid for the Academic Frontier Project, “Intelligent Information Science”, for Scientific Research on Innovative Areas (“Fluctuations and Biological Functions”), and for the Next Generation Super Computing Project, Nanoscience Program and Computational Materials Science Initiative from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

1. A. Liwo, C. Czaplewski, O. Stanislaw, H.A. Scheraga, *Curr. Opin. Struct. Biol.* **18**, 134 (2008)
2. H.A. Scheraga, *Ann. Rev. Biophys.* **40**, 1 (2011)
3. U.H.E. Hansmann, Y. Okamoto, *Curr. Opin. Struct. Biol.* **9**, 177 (1999)
4. A. Mitsutake, Y. Sugita, Y. Okamoto, *Biopolymers* **60**, 96 (2001)
5. Y. Okamoto, *J. Mol. Graphics Modell.* **22**, 425 (2004)
6. A. Mitsutake, Y. Mori, Y. Okamoto, in *Biomolecular Simulations: Methods and Protocols*, ed. by L. Monticelli, E. Salonen (Humana Press, Berlin, 2012), in press.
7. W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, J. Kenneth M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, *J. Am. Chem. Soc.* **117**, 5179 (1995)
8. P.A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, A. Pohorille, in *Computer Simulations of Biological Systems*, vol. 3, ed. by W.F. van Gunsteren, P.K. Weiner, A.J. Wilkinson (Kluwer/ESCOM, Dordrecht, 1997), pp. 83–96
9. J. Wang, P. Cieplak, P.A. Kollman, *J. Comput. Chem.* **21**, 1049 (2000)
10. V. Hornak, A. Abel, R. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **65**, 712 (2006)
11. Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, *J. Comput. Chem.* **24**, 1999 (2003)

12. A.D. MacKerell Jr, D. Bashford, M. Bellott, J. Dunbrack, R. L., J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, I. Reiher, W. E., B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, *J Phys Chem B* **102**, 3586 (1998)
13. A. MacKerell Jr, M. Feig, C. Brooks III, *J. Comput. Chem.* **25**, 1400 (2004)
14. A. MacKerell Jr, M. Feig, C. Brooks III, *J. Am. Chem. Soc.* **126**, 698 (2004)
15. W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225 (1996)
16. G.A. Kaminski, R.A. Friesner, J. Tirado-Rives, W.L. Jorgensen, *J. Phys. Chem. B* **105**, 6474 (2001)
17. W.F. Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hünenberger, P. Krüger, A.E. Mark, W.R.P. Scott, I.G. Tironi, (Vdf Hochschulverlag AG an der ETH Zürich, Zürich, 1996)
18. C. Oostenbrink, A. Villa, A.E. Mark, W.F.v. Gunsteren, *J. Comput. Chem.* **25**, 1656 (2004)
19. H.J.C. Berendsen, D. van der Spoel, R. van Drunen, *Comput. Phys. Commun.* **91**, 43 (1995)
20. E. Lindahl, B. Hess, D. van der Spoel, *J. Mol. Model.* **7**, 306 (2001)
21. G. Némethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H.A. Scheraga, *J. Phys. Chem.* **96**, 6472 (1992)
22. Y.A. Arnautova, A. Jagielska, H.A. Scheraga, *J. Phys. Chem. B* **110**, 5025 (2006)
23. T. Yoda, Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **386**, 460 (2004)
24. T. Yoda, Y. Sugita, Y. Okamoto, *Chem. Phys.* **307**, 269 (2004)
25. Y. Sakae, Y. Okamoto, *Chem. Phys. Lett.* **382**, 626 (2003)
26. Y. Sakae, Y. Okamoto, *J. Theo. Comput. Chem.* **3**, 339 (2004)
27. Y. Sakae, Y. Okamoto, *J. Theo. Comput. Chem.* **3**, 359 (2004)
28. C. Simmerling, B. Strockbine, A.E. Roitberg, *J. Am. Chem. Soc.* **124**, 11258 (2002)
29. Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, *J. Comput. Chem.* **24**, 1999 (2003)
30. M. Iwaoka, S. Tomoda, *J. Comput. Chem.* **24**, 1192 (2003)
31. N. Kamiya, Y. Watanabe, S. Ono, J. Higo, *Chem. Phys. Lett.* **401**, 312 (2005)
32. R.B. Best, G. Hummer, *J. Phys. Chem. B* **113**, 9004 (2009)
33. J. Mittal, R.B. Best, *Biophys. J.* **99**, L26 (2010)
34. Y. Sakae, Y. Okamoto, *J. Phys. Soc. Jpn.* **75** (2006). 054802 (9 pages)
35. Y. Sakae, Y. Okamoto, *Mol. Sim.* **36**, 138 (2010)
36. G.N. Ramachandran, V. Sasisekharan, *Adv. Protein Chem.* **23**, 283 (1968)
37. Y. Sakae, Y. Okamoto, *Mol. Sim.* **36**, 159 (2010)
38. Y. Sakae, Y. Okamoto, *Mol. Sim.* **36**, 1148 (2010)
39. Y. Sakae, Y. Okamoto, e-print: arXiv:1206.3909 [cond-mat.stat-mech]; submitted for publication.
40. Y. Sakae, Y. Okamoto, *Mol. Sim.* In press.
41. W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127 (1990)
42. D. Qiu, P.S. Shenkin, F.P. Hollinger, W.C. Still, *J. Phys. Chem. A* **101**, 3005 (1990)
43. S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, *Science* **220**, 671 (1983)
44. W. Kabsch, C. Sander, *Biopolymers* **22**, 2577 (1983)
45. Y. Sakae, Y. Okamoto, In preparation.
46. S. Honda, N. Kobayashi, E. Munekata, *J. Mol. Biol.* **295**, 269 (2000)
47. K.R. Shoemaker, P.S. Kim, D.N. Brems, S. Marqusee, E.J. York, I.M. Chaiken, J.M. Stewart, R.L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2349 (1985)
48. J.J. Osterhout Jr., R.L. Baldwin, E.J. York, J.M. Stewart, H.J. Dyson, P.E. Wright, *Biochemistry* **28**, 7059 (1989)
49. F.J. Blanco, G. Rivas, L. Serrano, *Nature Struct. Biol.* **1**, 584 (1994)
50. N. Kobayashi, S. Honda, H. Yoshii, H. Uedaira, E. Munekata, *FEBS Lett.* **366**, 99 (1995)
51. Accelrys discovery studio visualizer. Software available at <http://www.accelrys.com/>
52. H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, *J. Chem. Phys.* **81**, 3684 (1984)
53. Tinker program package. Software available at <http://dasher.wustl.edu/tinker/>
54. URL <http://www.accelrys.com/>

- 55. T. Noguchi, K. Onizuka, Y. Akiyama, M. Saito, in *Proc. of the Fifth International Conference on Intelligent Systems for Molecular Biology* (AAAI press, Menlo Park, CA, 1997)
- 56. D.A. Case, T. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, R. Woods, *J. Computat. Chem.* **26**, 1668 (2005)
- 57. A. Onufriev, D. Bashford, D.A. Case, *Proteins* **55**, 383 (2004)
- 58. J. Weiser, P.S. Shenkin, W.C. Still, *J. Comput. Chem.* **20**, 217 (1999)
- 59. Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **314**, 141 (1999)
- 60. J.P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, *J. Comput. Phys.* **23**, 327 (1977)
- 61. G. Wang, R.L.D. Jr, *Bioinformatics* **19**, 1589 (2003)
- 62. W.G. Hoover, *Phys. Rev. A* **31**, 1695 (1985)
- 63. W.L. Jorgensen, J. Tirado-Rives, *J. Am. Chem. Soc.* **110**, 1657 (1988)
- 64. M. Levitt, C. Chothia, *Nature* **261**, 552 (1976)